SUPERMARKET SALES ANALYSIS

DATA SET EXPLORATION PART 3

Moganaviniith Rathinavel

11/22/13

BDAT 1005 – 22F
by Jonathan Gladstone

Appropriate Data Set Description

Data set description has already listed in previous data set exploration file and there is no need to change it.

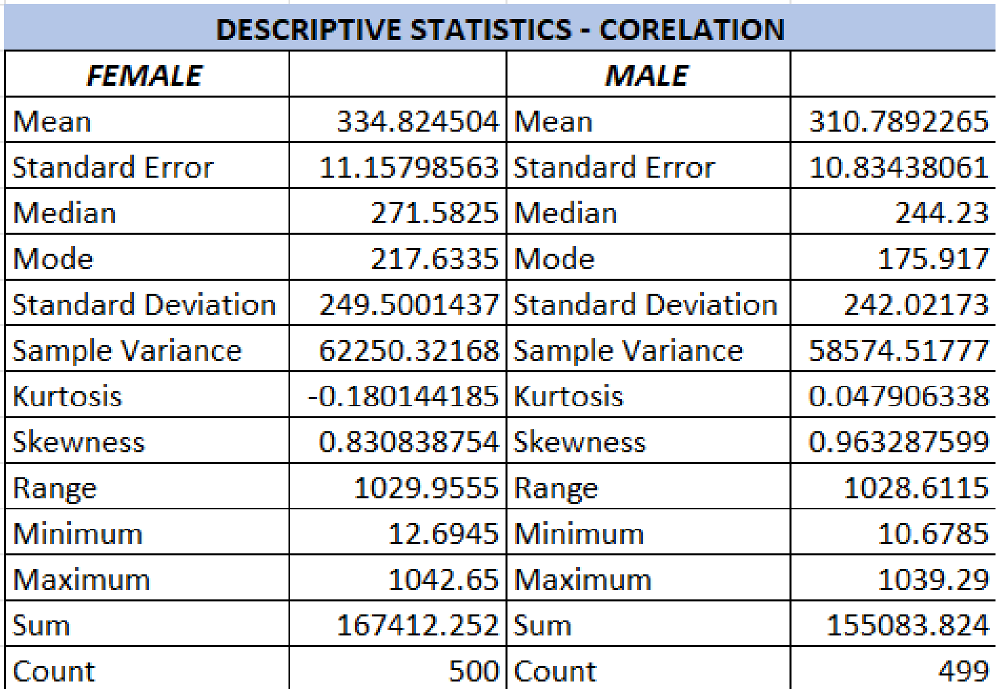
Univariate descriptive statistics

As same as last submitted document.

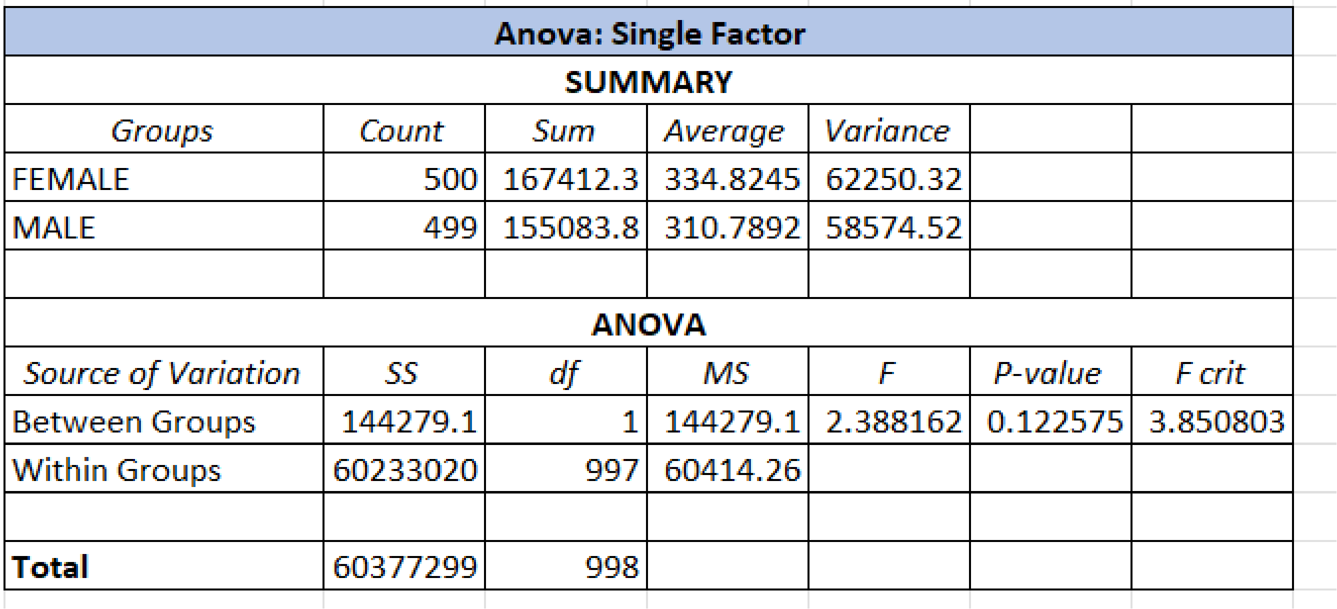
Hypothesis and Tests

1.Anova test of male and female total:

- 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 one way test and 2nd is 2 way test.
- Here we have used the anova with single factor for our hypothesis.

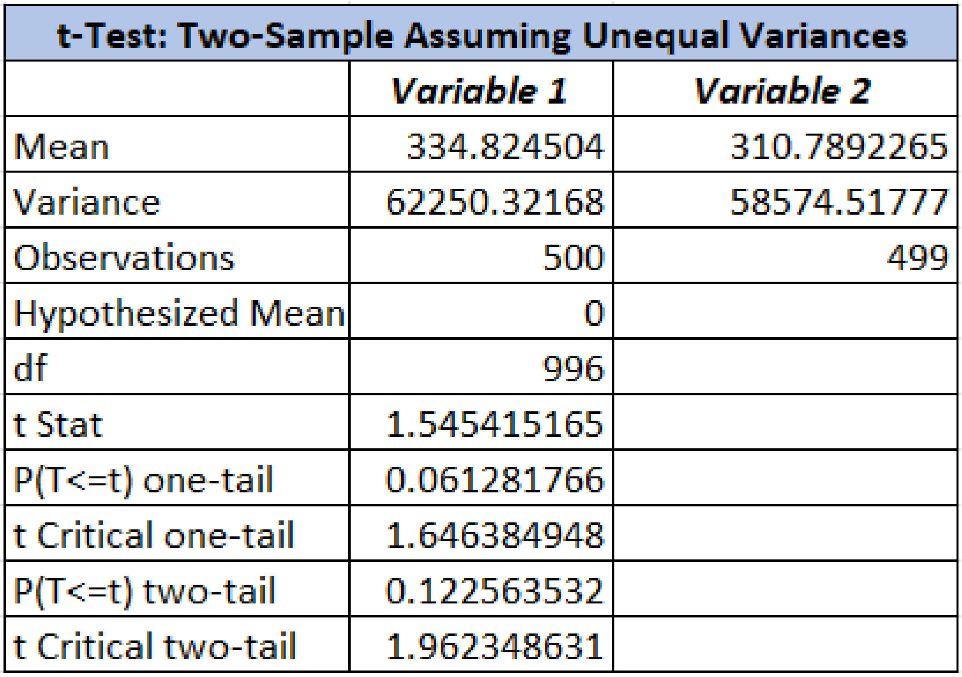


DESCRIPTIVE STATISTICS - CORELATION			
FEMALE		***MALE***	
Mean	334.824504	Mean	310.7892265
Standard Error	11.15798563	Standard Error	10.83438061
Median	271.5825	Median	244.23
Mode	217.6335	Mode	175.917
Standard Deviation	249.5001437	Standard Deviation	242.02173
Sample Variance	62250.32168	Sample Variance	58574.51777
Kurtosis	-0.180144185	Kurtosis	0.047906338
Skewness	0.830838754	Skewness	0.963287599
Range	1029.9555	Range	1028.6115
Minimum	12.6945	Minimum	10.6785
Maximum	1042.65	Maximum	1039.29
Sum	167412.252	Sum	155083.824
Count	500	Count	499

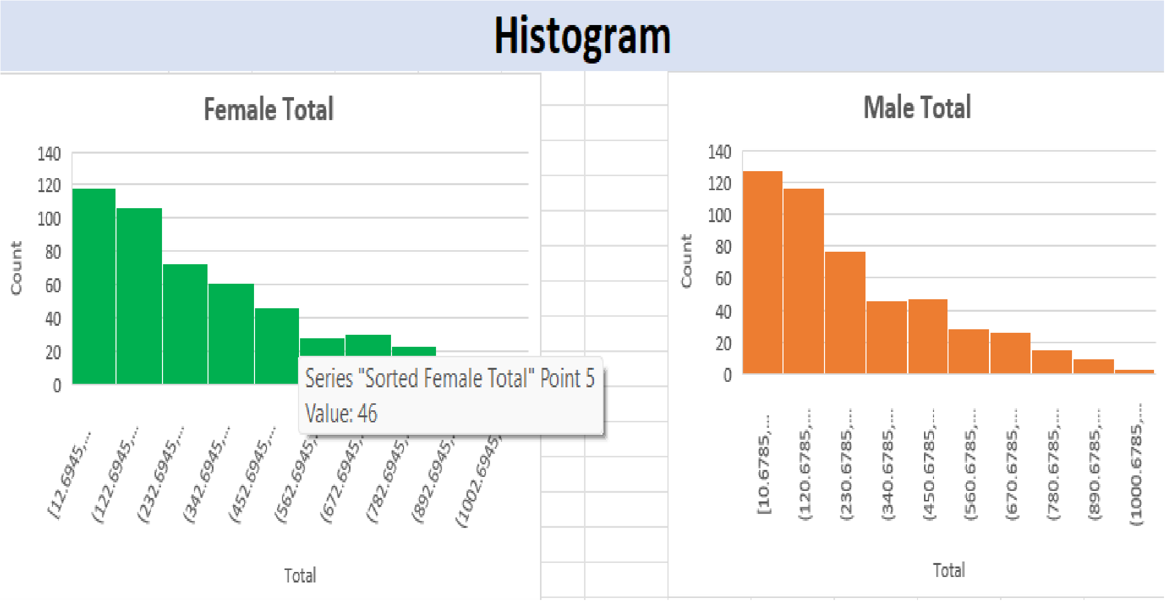


Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
FEMALE	500	167412.3	334.8245	62250.32		
MALE	499	155083.8	310.7892	58574.52		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	144279.1	1	144279.1	2.388162	0.122575	3.850803
Within Groups	60233020	997	60414.26			
Total	60377299	998				

Also performed t-test between 2 variables.

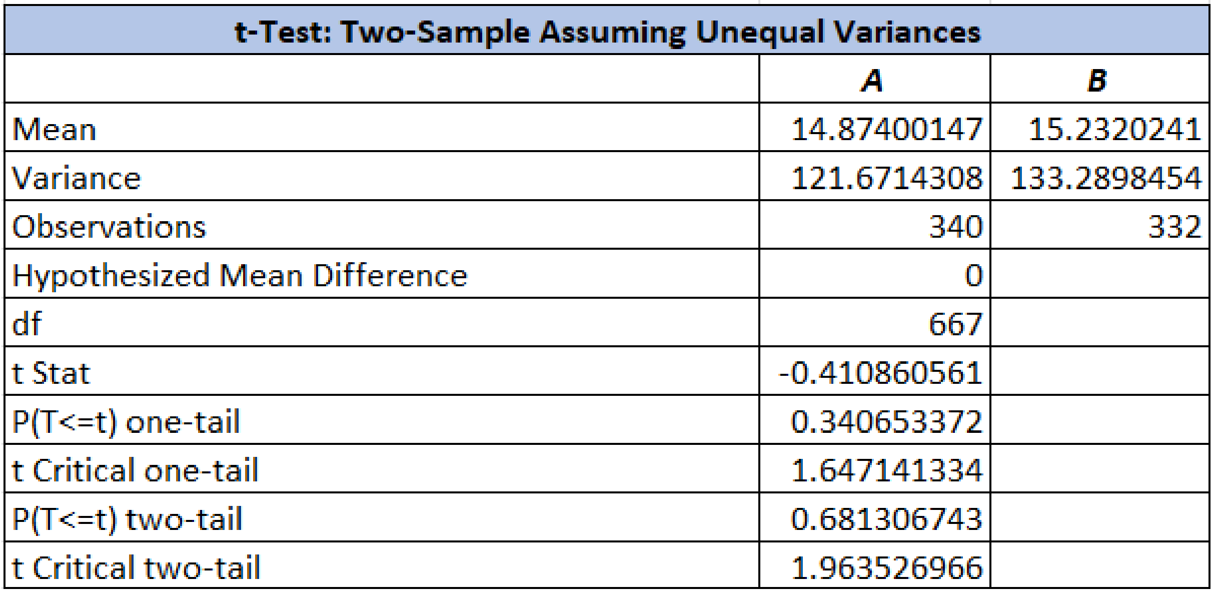


t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	***Variable 2***
Mean	334.824504	310.7892265
Variance	62250.32168	58574.51777
Observations	500	499
Hypothesized Mean	0	
df	996	
t Stat	1.545415165	
P(T<=t) one-tail	0.061281766	
t Critical one-tail	1.646384948	
P(T<=t) two-tail	0.122563532	
t Critical two-tail	1.962348631	



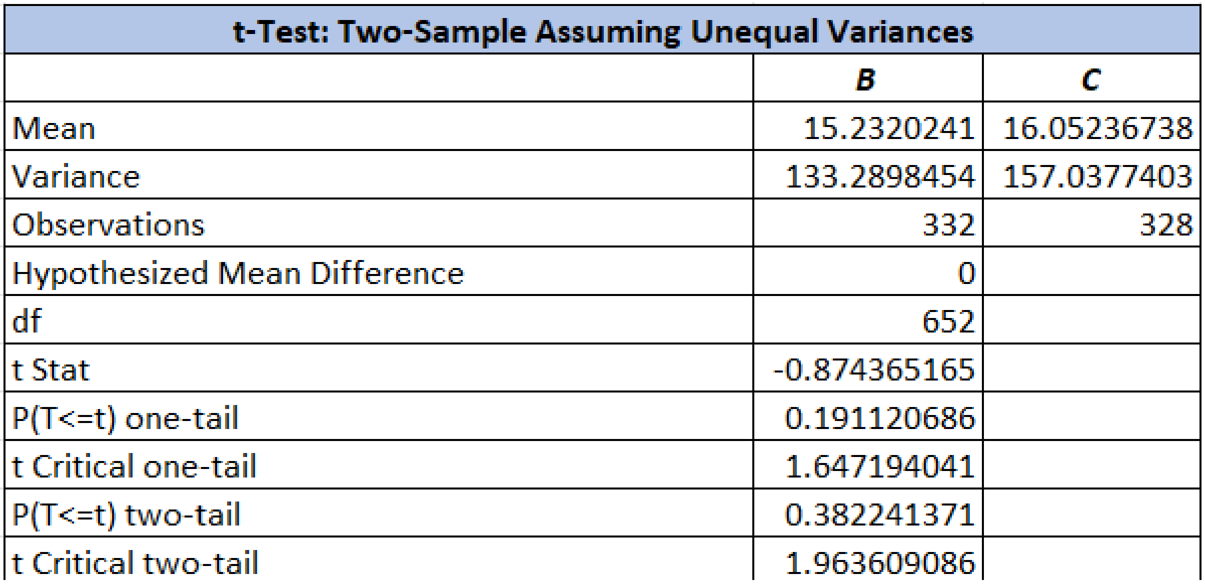
2. T test between branches A,B and C

- T test between A & B



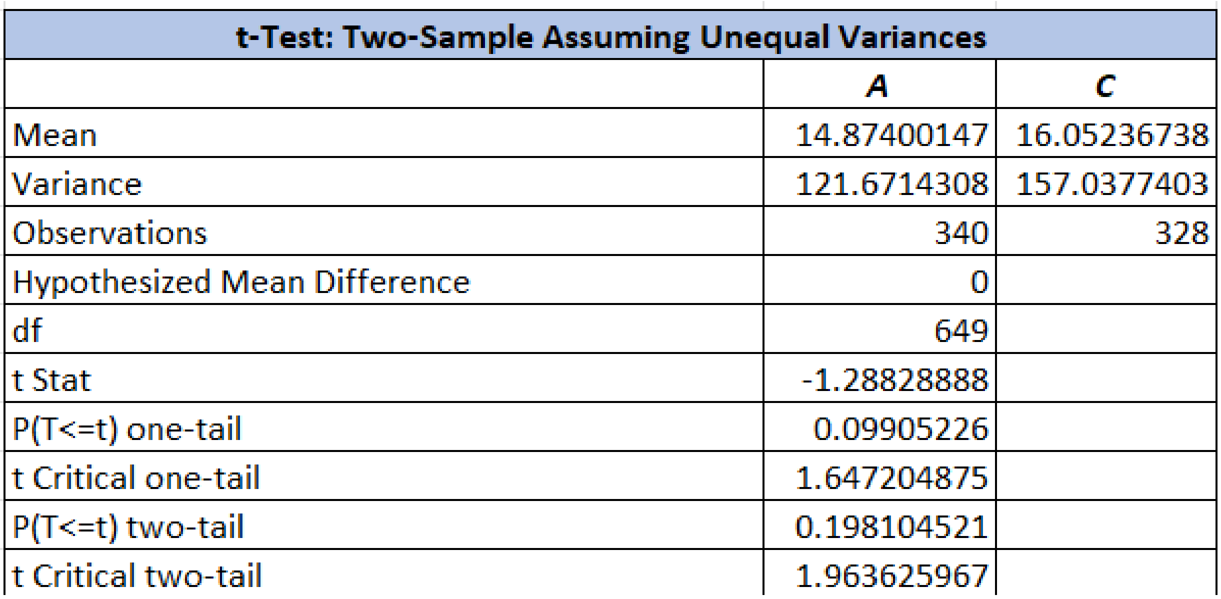
t-Test: Two-Sample Assuming Unequal Variances		
	A	*B*
Mean	14.87400147	15.2320241
Variance	121.6714308	133.2898454
Observations	340	332
Hypothesized Mean Difference	0	
df	667	
t Stat	-0.410860561	
P(T<=t) one-tail	0.340653372	
t Critical one-tail	1.647141334	
P(T<=t) two-tail	0.681306743	
t Critical two-tail	1.963526966	

- T test between B & C



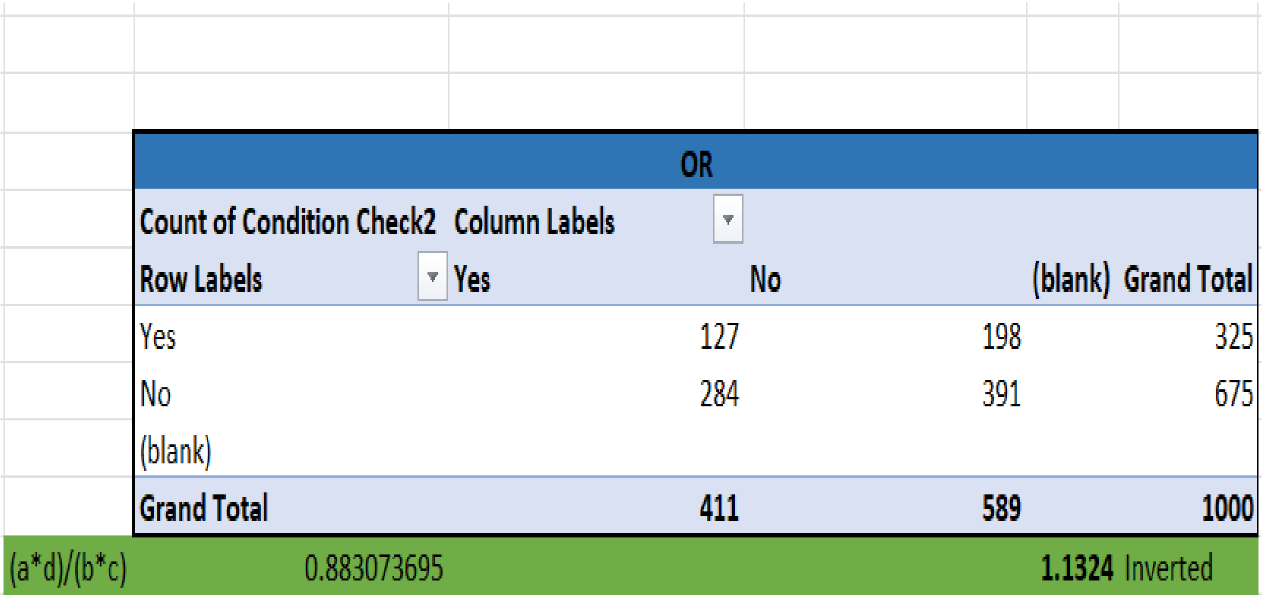
t-Test: Two-Sample Assuming Unequal Variances		
	B	***C***
Mean	15.2320241	16.05236738
Variance	133.2898454	157.0377403
Observations	332	328
Hypothesized Mean Difference	0	
df	652	
t Stat	-0.874365165	
P(T<=t) one-tail	0.191120686	
t Critical one-tail	1.647194041	
P(T<=t) two-tail	0.382241371	
t Critical two-tail	1.963609086	

- T test of A & C



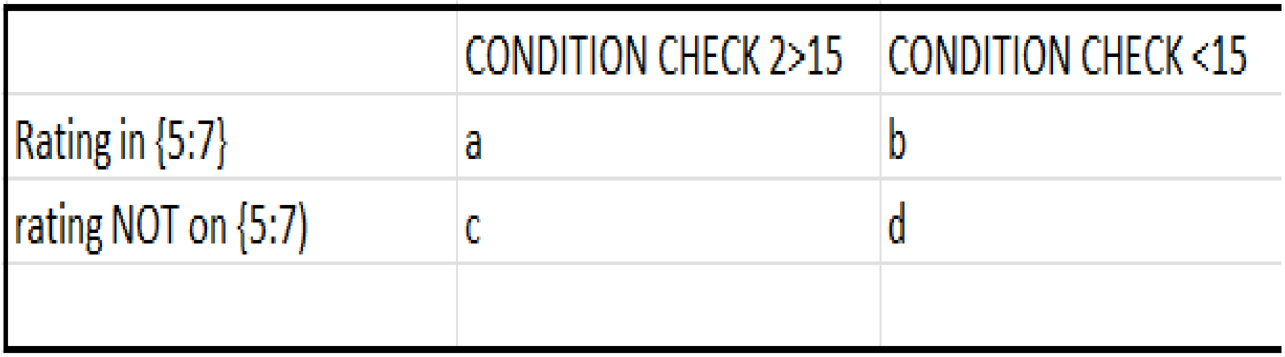
t-Test: Two-Sample Assuming Unequal Variances		
	A	***C***
Mean	14.87400147	16.05236738
Variance	121.6714308	157.0377403
Observations	340	328
Hypothesized Mean Difference	0	
df	649	
t Stat	-1.28828888	
P(T<=t) one-tail	0.09905226	
t Critical one-tail	1.647204875	
P(T<=t) two-tail	0.198104521	
t Critical two-tail	1.963625967	

3. Odd Risk test of rating and gross income



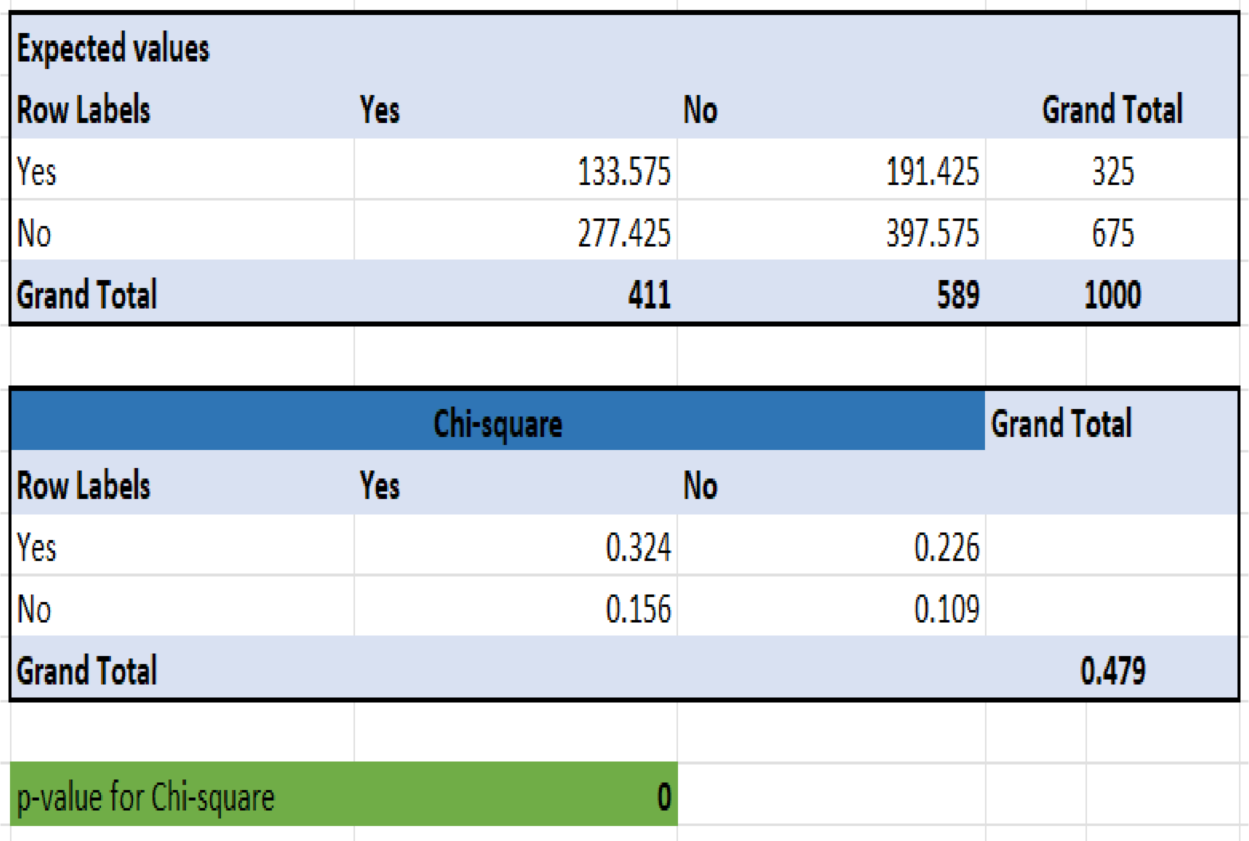
	OR				
	Count of Condition Check2	**Column Labels** ▾			
	Row Labels ▾	**Yes**	**No**	**(blank)**	**Grand Total**
	Yes	127	198		325
	No	284	391		675
	(blank)				
	Grand Total	**411**	**589**		**1000**
(a*d)/(b*c)	0.883073695			**1.1324**	Inverted

Condition check: (after setting range)



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

4. Chi-Square test



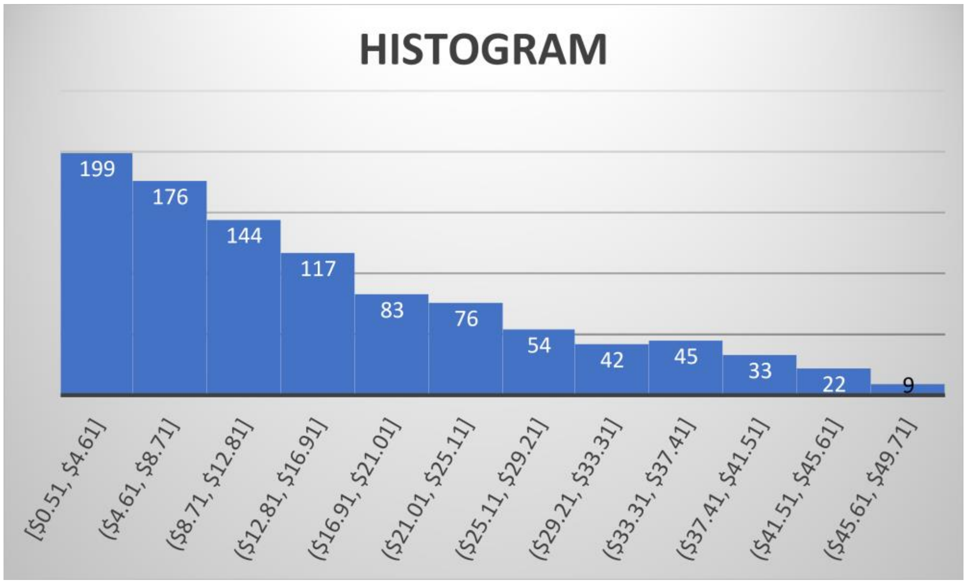
Expected values

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**

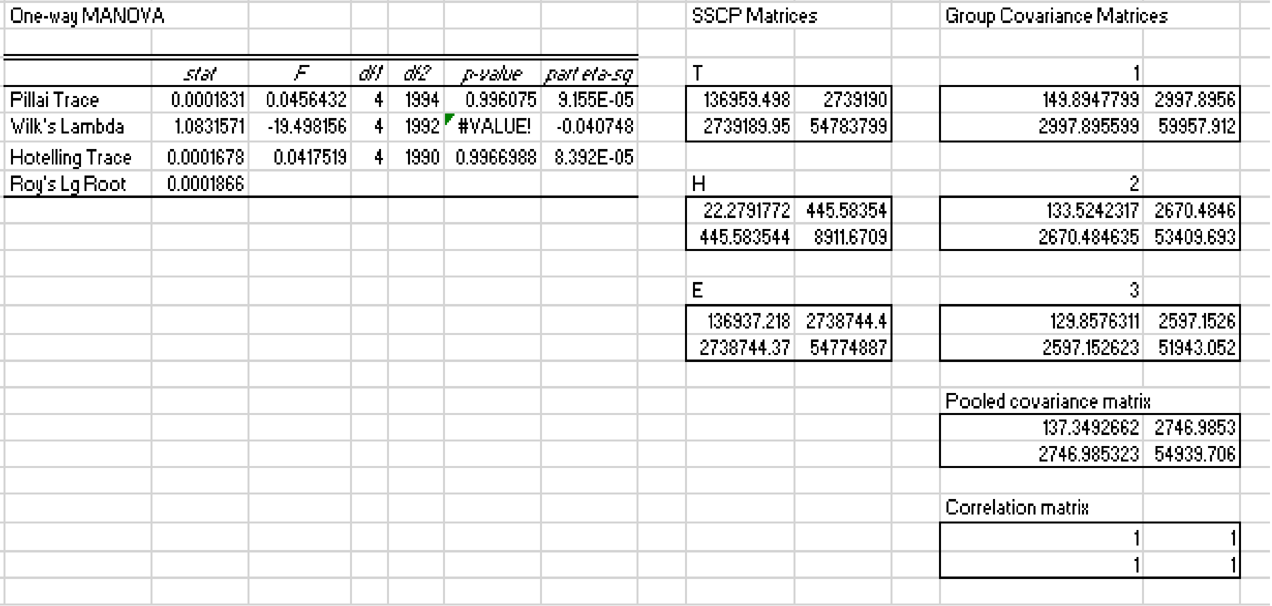
Histogram of rating and gross income



5. Manova

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

Used invoice id, sum of gross income, sum of cogs and coded variable.



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

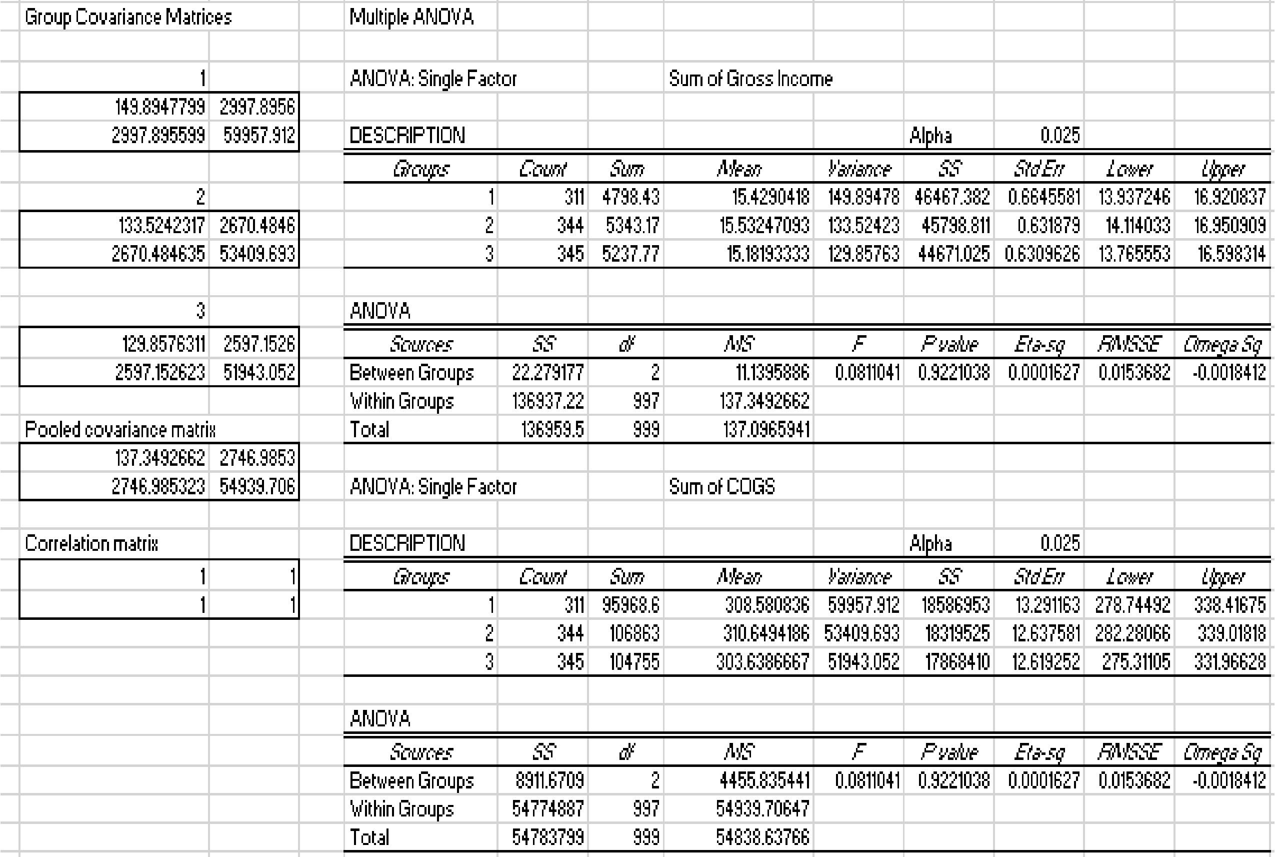
129.8576311	2597.1526
2597.152623	51943.052

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					

Finer Research Question

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?
    - Generally, ratings not only depend on customer service one gets, but also depend on the price of the goods as customer may compare the price from one store with another. And using the current trend future gross income may be calculated
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
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?

Demonstrated Tracking

First of all, we started with developing hypothesis test. Then continued with performing many tests such as:

1. Anova test using female, male variable’s total, included co-relation and t-test.
2. T-test on branches A,B,C.
3. Odd risk test of rating and gross income.
4. Chi-square test
5. 5. Manova.

References

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