Group Project Report

Group 19

Thulari Jayasinghe - s15540

Erantha Pingewatta - s15549

Chamani Pramoda - s15551

Supipi Senarath - s15563

Contents

1.0 Introduction

The “Supermarket sales” data set includes information about sales of supermarket company which has recorded in 3 different branches A, B, C, and the supermarket sales data set is a census of sales data of supermarket which has 1000 customers.

The variables in this data set are, Invoice Id, Branch, City, C_type, Gender, Product line, Unit price, Quantity, Tax, Total, Payment, cogs, gross_perc, gross_income, and Rating.

Invoice id: Computer generated sales slip invoice identification number
Branch: Branch of supercenter (3 branches are available identified by A, B and C).
City: Location of supercenters
C_ type: Type of customers, recorded by Members for customers using member card and Normal for without member card.
Gender: Gender type of customer
Product line: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel
Unit price: Price of each product in $
Quantity: Number of products purchased by customer
Tax: 5% tax fee for customer buying
Total: Total price including tax
Payment: Payment used by customer for purchase (3 methods are available – Cash, Credit card and EWallet)
COGS: Cost of goods sold
Gross_perc: Gross margin percentage
Gross_income: Gross income
Rating: Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

Dataset was analyzed using Simple Random Sampling, Stratified Sampling, and Cluster Sampling separately. All these methods are explained in detail in the next parts of the report.

2.0 Methodology

Sample size calculation

Simple Random Sampling is a type of probability sampling technique in which every possible subset of n distinct units in the population has the same probability of being selected as the sample.

When doing a survey, this equation is used to calculate the sample size.

$$\mathbf{n} = \frac{n0}{1+(\frac{n0}{N})} \qquad \text{where} \qquad n0 = \left(Z\,\alpha/2\,\frac{s}{e}\right)^2$$

n = Sample Size

N = Population size

$Z_{\alpha/2}$ = Z value of the confidence level 1- α

S = Population standard deviation

e = Margin of error

1. Simple Random Sampling

The analysis was done using the R software package. For the analysis 2 samples were selected from the population

Sample size was calculated using the above equations with a margin of error of 3. The value got is 517 The “rsamp” function included in the “sampler” package in R software is used to extract two simple random samples of size 517 from the population.

2. Stratified Random Sampling

The analysis was done using the R software package. For the analysis 2 samples were selected from the population. A random sample size was generated using “rsampcalc” command with tolerable margin of error of 3. The value got is 517. Then by using simple random sampling method, samples were selected from each stratum according to weights to get the total sample size of 517. For that “ssampcalc” command is used.

Stratification variable

To do stratify, first the population of N sampling units is divided into H “layers” or strata, with N_h sampling units in stratum h. The values of $N_1, N_2, \ldots, N_H$ should be known.

$$N_1 + N_2 + \cdots + N_H = N$$

In stratified random sampling an SRS is taken independently from each stratum, so that n_h observations are randomly selected from the Nh population units in stratum h. S_h is defined to be the set of n_h units in the SRS for stratum h. The total sample size is,

$$n = n_1 + n_2 + \cdots + n_H$$

To divide the population into H strata, a stratification variable should be selected.

When selecting the stratification variable, it also considered that strata do not overlap and the variance within the strata are minimum.

We did stratification using the variable product line since we expect customers will spend around the same amount in a particular product category. (i.e., people will spend roughly the same for food) and among product categories the expenditures are expected to have some more variation (i.e., contrast between electronic expenses vs health, travel vs fashion) This variable had 6 levels in it. Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports, and Travel. So, 6 stratums were obtained.

	Product.line	Nh	wt[,1]	nh[,1]
	<chr>	<int>	<dbl>	<dbl>
1	Electronic accessories	170	0.17	88
2	Fashion accessories	178	0.178	92
3	Food and beverages	174	0.174	90
4	Health and beauty	152	0.152	79
5	Home and lifestyle	160	0.16	83
6	Sports and travel	166	0.166	86

>

3. Two stage cluster Sampling

In contrast to SRS and stratified sampling techniques where we need a sampling frame, we choose for cluster sampling technique when faced with difficulties creating a sampling frame and additionally it is more cost effective.

We divide the population into naturally occurring groups (geographical / organizational). Here we employed two stage cluster sampling technique.

We considered "City" as primary sampling unit as we can expect similar variance among the Cities with respect to the variable of interest (supermarket sales) Thereby satisfying our need to make clusters similar as possible.

Out of the 3 cities:

```
Mandalay Naypyitaw    Yangon
     332       328       340
```

We randomly selected 2 psus:

```
Mandalay Naypyitaw
     332       328
```

in the 1st stage of the cluster sampling procedure. Then in the second stage we took an SRS of each cluster selected with the sampling size required to maintain the margin of error at 3.

The sample was as following:

```
Mandalay Naypyitaw
     254       251
```

The sampling weights for Mandalay: 1.96063 while for Naypyitaw: 1.960159

3.0 Results of the Study

I: Simple Random Sampling

Population Data

Mean	rating	6.9727					
	total	322.9667					
	gross_income	15.379369					
Total	gross_income	15379.369					
	tax	15379.369					
	total_cost	307587.38					

Proportion	c_type	Member 0.501	Normal 0.499				
	payment	Cash 0.344	Credit card 0.311	Ewallet 0.345			
	gender	Female 0.501	Male 0.499				
	product_line	Electronic 0.170	Fashion 0.187	Food and beverages 0.174	Health and beauty 0.152	Home and lifestyle 0.160	Sports and travel 0.166

Sample Data

For sample 1

Mean	total	mean 334.88	SE 11.042	
	rating	mean 6.9025	SE 0.0754	
	gross_income	mean 15.947	SE 0.5258	
Total	tax	total 8244.5	SE 271.85	
	cost	total 164890	SE 5436.9	
	gross_income	total 8244.5	SE 271.85	
Proportion	customer_type	 Member Normal	Mean 0.51451 0.48549	SE 0.022 0.022
	pay_method	 Cash Credit Ewallet	Mean 0.32108 0.32302 0.35590	SE 0.0206 0.0206 0.0211
	gender	 Male female	mean 0..5087 0.4913	SE 0.022 0.022
	branch	 A B C	mean 0.33849 0.32689 0.33462	SE 0.0208 0.0206 0.0208
Ratio	Total/tax 5%	ratio 21		SE 7.231773

For sample 2

Mean	total	mean 322.84	SE 11.007	
	rating	mean 6.9576	SE 0.0752	
	gross_income	mean 15.374	SE 0.5242	
Total	tax	total 7948.1	SE 270.99	
	cost	total 158963	SE 5419.8	
	gross_income	total 7948.1	SE 270.99	
Proportion	customer_type	 Member Normal	Mean 0.51451 0.48549	SE 0.022 0.022
	pay_method	 Cash Credit Ewallet	Mean 0.35203 0.29207 0.35590	SE 0.0210 0.0200 0.0211
	gender	 Male female	mean 0..48936 0.51064	SE 0.022 0.022
	branch	 A B C	mean 0.35590 0.31915 0.32495	SE 0.0211 0.0205 0.0206
Ratio	Total/tax 5%	Ratio 21	SE 7.716507e-17	

Here, SRS 01 and SRS 02 are given nearly equivalent estimated values for all the variables we considered. When comparing the sample 1 and sample 2 estimations with the population values all three estimators mean, total and proportion are approximately equivalent with lower standard errors.

+It is justifiable to say the mean and proportion estimates are close to the population values. But the total estimates are different from the population values.

pop total (307587) vs srs1 total (164890) & srs2_total (158963)

pop_total_tax(15379) vs srs1_total-tax(8244) & srs2tot_tax(7948)

II: Stratified Sampling

Stratification Variable: Product line

Sample Size: 517

```
  Product.line               Nh wt[,1] nh[,1]
  <chr>                   <int>  <dbl>  <dbl>
1 Electronic accessories    170  0.17      88
2 Fashion accessories       178  0.178     92
3 Food and beverages        174  0.174     90
4 Health and beauty         152  0.152     79
5 Home and lifestyle        160  0.16      83
6 Sports and travel         166  0.166     86
> |
```

For sample 1

Mean	total	mean 324.86	SE 10.937	
	rating	mean 6.9725	SE 0.0755	
	gross_income	mean 15.47	SE 0.5208	
Total	total	total 28120.02	SE 946.68	
	gross_income	total 1339.05	SE 45.08	
	cogs	total 26781.0	SE 901.6	
	tax	Total 1339.05	SE 45.08	
Proportion	customer_type	 Member Normal	Mean 0.47542 0.52458	SE 0.022 0.022
	pay_method	 Cash Credit Ewallet	Mean 0.32320 0.30890 0.36791	SE 0.0205 0.0203 0.0212

	gender	 Male female	mean 0..51814 0.48186	SE 0.0219 0.0219
	branch	 A B C	mean 0.32657 0.34857 0.32486	SE 0.0207 0.0211 0.0206
Ratio	Total/tax 5%	Ratio 21	SE 7.660147e-17	

R 4.3.1 · D:/Campus/Level 3/Sem 1/IS 3001/Grp.Proj/Project/

```
> #mean rating by strata
> svyby(~Rating,by=~Product.line,strat1_des,svymean)
                                 Product.line   Rating        se
Electronic accessories Electronic accessories 6.871591 0.1772301
Fashion accessories       Fashion accessories 6.906522 0.1900902
Food and beverages         Food and beverages 7.185556 0.1750374
Health and beauty           Health and beauty 7.093671 0.1942649
Home and lifestyle         Home and lifestyle 6.872289 0.1832969
Sports and travel           Sports and travel 6.911628 0.1866504
> #mean total by strata
> svyby(~Total,by=~Product.line,strat1_des,svymean)
                                 Product.line    Total       se
Electronic accessories Electronic accessories 315.5424 27.51963
Fashion accessories       Fashion accessories 295.5458 25.38116
Food and beverages         Food and beverages 321.2064 25.59268
Health and beauty           Health and beauty 321.4320 26.86406
Home and lifestyle         Home and lifestyle 329.6962 27.33900
Sports and travel           Sports and travel 370.6524 27.79516
>
> #total tax by strat
> svyby(~Tax.5.,by=~Product.line,strat1_des,svytotal)
                                 Product.line   Tax.5.       se
Electronic accessories Electronic accessories 224.7864 19.60446
Fashion accessories       Fashion accessories 230.4694 19.79247
Food and beverages         Food and beverages 239.5282 19.08483
Health and beauty           Health and beauty 183.7979 15.36113
Home and lifestyle         Home and lifestyle 208.4936 17.28866
Sports and travel           Sports and travel 251.9731 18.89541
>
```

For sample 2

Mean	total	mean 328.5	SE 11.105	
	rating	mean 7.063	SE 0.0755	
	gross_income	mean 15.643	SE 0.5288	
Total	total	total 28434.72	SE 961.26	
	gross_income	total 1354.034	SE 45.774	
	cogs	total 27080.68	SE 915.49	
	tax	Total 1354.034	SE 45.774	
Proportion	customer_type	 Member Normal	Mean 0.48764 0.51236	SE 0.0219 0.0219
	pay_method	 Cash Credit Ewallet	Mean 0.35527 0.30381 0.34092	SE 0.0211 0.0203 0.0218
	gender	 Male female	mean 0.48348 0.51652	SE 0.0219 0.0219
	branch	 A B C	mean 0.35663 0.32059 0.32278	SE 0.0211 0.0206 0.0206
Ratio	Total/tax 5%	Ratio 21	SE 7.660147e-17	

```
Console   Terminal   Background Jobs
R 4.3.1 · D:/Campus/Level 3/Sem 1/IS 3001/Grp.Proj/Project/
> #>>>>>>>>>>>sub groups
> #mean rating by strata
> svyby(~Rating,by=~Product.line,strat2_des,svymean)
                                  Product.line   Rating        se
Electronic accessories Electronic accessories 7.022727 0.1699755
Fashion accessories       Fashion accessories 7.222826 0.1889386
Food and beverages         Food and beverages 7.197778 0.1798250
Health and beauty           Health and beauty 6.869620 0.2024305
Home and lifestyle         Home and lifestyle 6.960241 0.1773622
Sports and travel           Sports and travel 7.032558 0.1895451
> #mean total by strata
> svyby(~Total,by=~Product.line,strat2_des,svymean)
                                  Product.line    Total       se
Electronic accessories Electronic accessories 337.4136 27.06836
Fashion accessories       Fashion accessories 294.8313 25.49443
Food and beverages         Food and beverages 343.7101 26.32676
Health and beauty           Health and beauty 318.4873 27.36113
Home and lifestyle         Home and lifestyle 335.7746 27.76114
Sports and travel           Sports and travel 342.7338 29.04304
>
> #total tax by strat
> svyby(~Tax.5.,by=~Product.line,strat2_des,svytotal)
                                  Product.line   Tax.5.       se
Electronic accessories Electronic accessories 240.3670 19.28299
Fashion accessories       Fashion accessories 229.9123 19.88080
Food and beverages         Food and beverages 256.3096 19.63224
Health and beauty           Health and beauty 182.1141 15.64535
Home and lifestyle         Home and lifestyle 212.3374 17.55562
Sports and travel           Sports and travel 232.9937 19.74373
>
```

Both samples give roughly the same values for the parameter estimates, especially with respect to mean and proportion estimates. The SEs are almost the same for every estimate. The estimates have lower SE although estimates for total have comparatively higher SE. It may be the reason that the stratums we considered with respect to stratification variable are not actually different (i.e.: the variation among strata is low) It is obvious in distribution of means and totals of rating, Total and Tax variables (SEs and the estimates seems similar)

Estimates from the 2 samples seems to go with the population values but the estimates for totals are way off example: total cost_pop = 307587 >>> 26781 &27080, total tax_pop = 15379 >> 1339 & 1354 etc.

III: Cluster Sampling

Clustering Variable: City

For sample 1

Mean	total	mean 323.6408	SE 8.1663	
	rating	mean 7.034473	SE 0.024834	
	gross_income	mean 15.41147	SE 0.38887	
Total	tax	total 15442	SE 112.24	
	cost	total 308846	SE 2244.8	
	gross_income	total 15442	SE 112.24	
Proportion	customer_type	 Member Normal	Mean 0.51089 0.48911	SE 0.003 0.003
	pay_method	 Cash Credit Ewallet	Mean 0.34059 0.31683 0.34257	SE 0.0139 0.0140 0.0001
	gender	 Male female	mean 0..50693 0.49307	SE 0.0149 0.0149
	branch	 B C	mean 0.32689 0.33462	SE 0.5 0.5

For sample 2

Mean	total	mean 324.3282	SE 1.3033	
	rating	mean 6.8768	SE 0.1161	
	gross_income	mean 15.444199	SE 0.062063	
Total	tax	total 15290	SE 154.11	

	cost	total 305795	SE 3082.2	
	gross_income	total 15290	SE 154.11	
Proportion	customer_type	 Member Normal	Mean 0.493 0.507	SE 0.0284 0.0284
	pay_method	 Cash Credit Ewallet	Mean 0.35744 0.31243 0.33012	SE 0.0285 0.0134 0.0151
	gender	 Male female	mean 0..50668 0.49332	SE 0.0463 0.0463
	branch	 A C	mean 0.50898 0.49102	SE 0.4998 0.4998

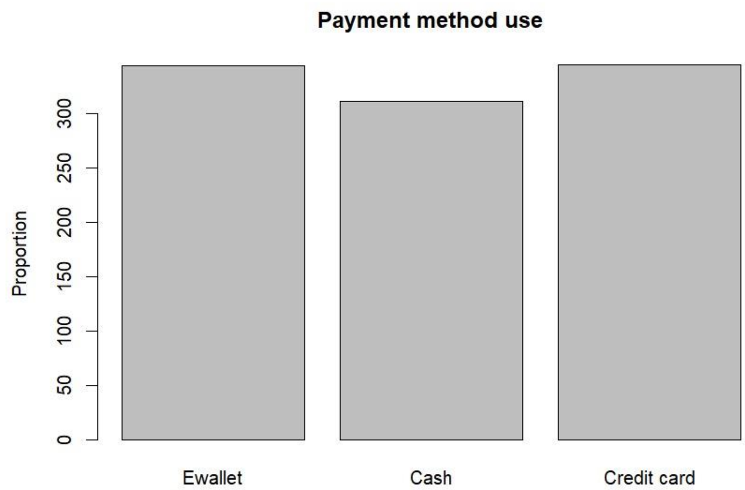
When studying the estimations from the cluster sample the estimated mean, proportion, total and SE are different to each other. Take mean of the total of the sample 1 is less than the mean of the total of the sample 2 but SE of the mean of the total of the sample 1 is greater than the mean of the total of the sample 2.

Now consider total of the tax variable, in here total of the tax is slightly equal in the two samples but the SE are different. SE of the total of the tax of sample 2 is greater than sample 1.

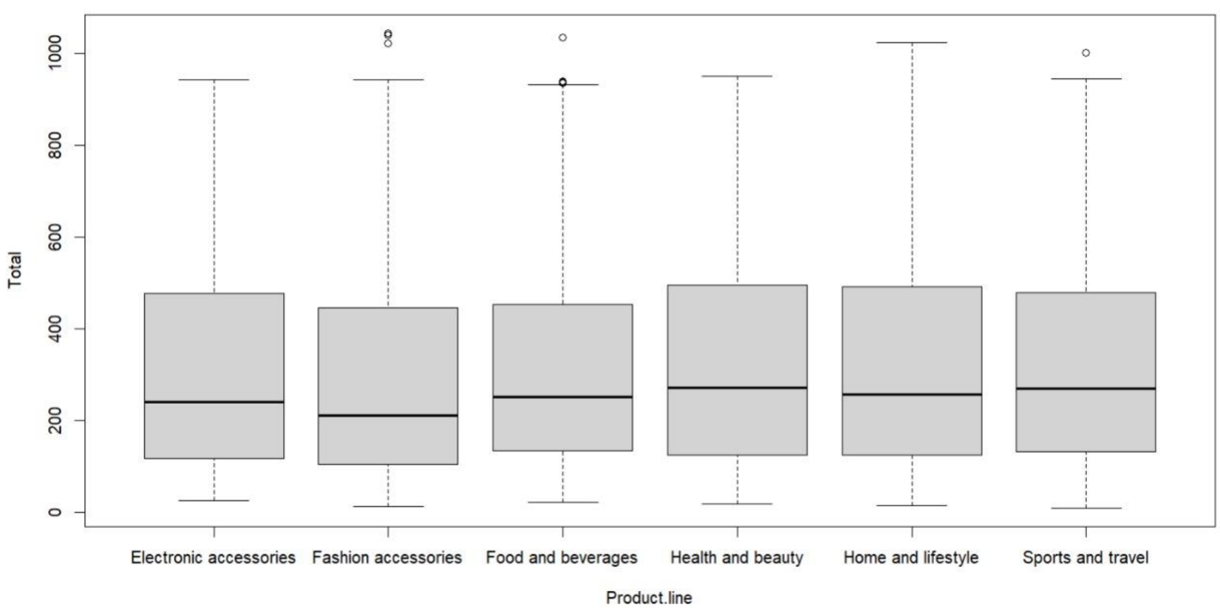
When considering proportion of the customer type of the two samples are approximately equal but SE of the proportion of the customer type of sample 2 is greater than the sample 1.

So, we can conclude that it might be the case that the clusters we considered do not have similar variation among each other. Clusters are different from each other.

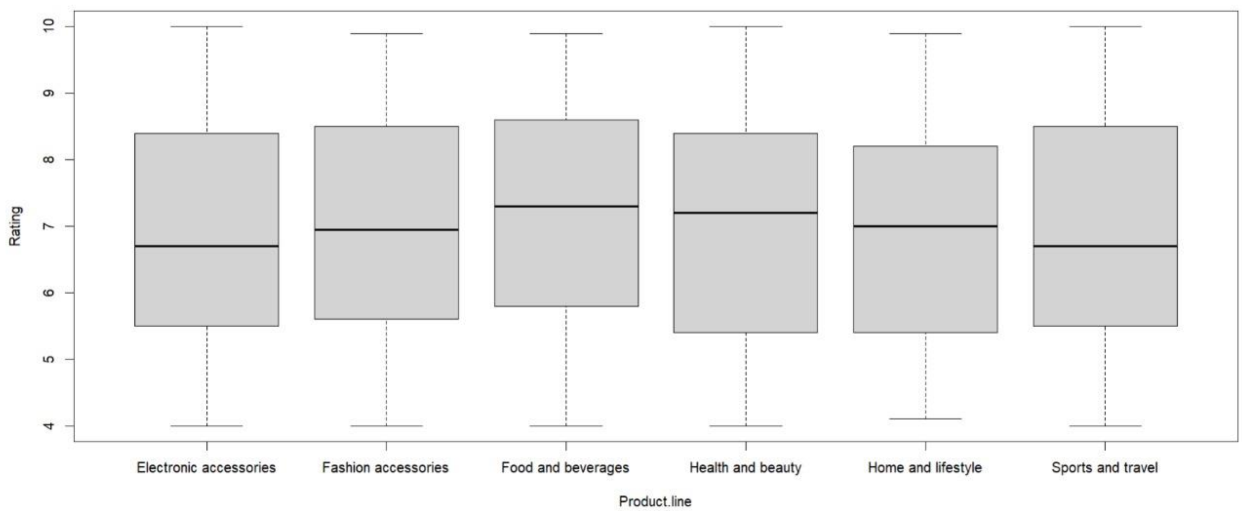
4.0 Graphical Analysis



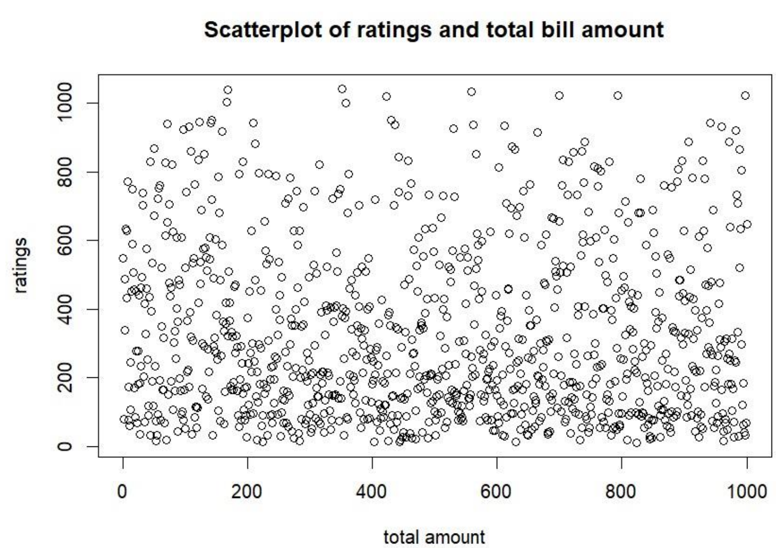
The above graph compares the proportion values of each payment method. There are EWallet, Cash and Credit card. The Most famous methods are Ewallet and credit card. They have slightly equal values. When compared to the other payments methods number of people who has used cash is low.



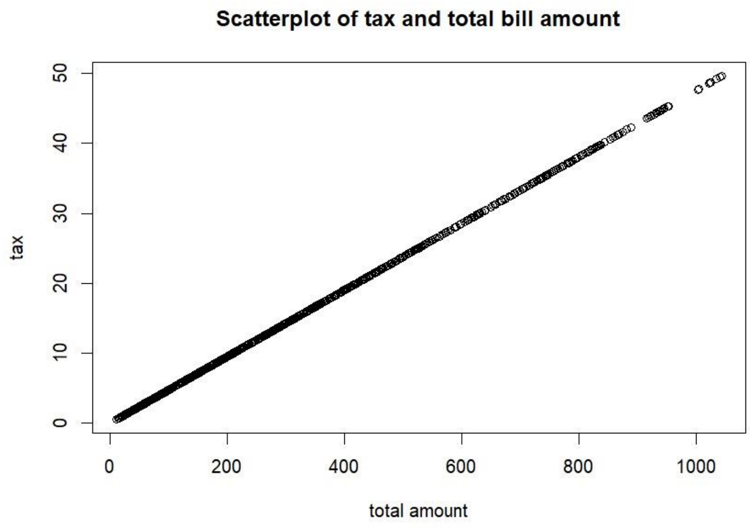
The above box plot compares the value of total and product line. We can observe that some outliers are in the fashion accessories, food and beverages and sports and travel. Seems like health and beauty and sports and travels have the highest mean.



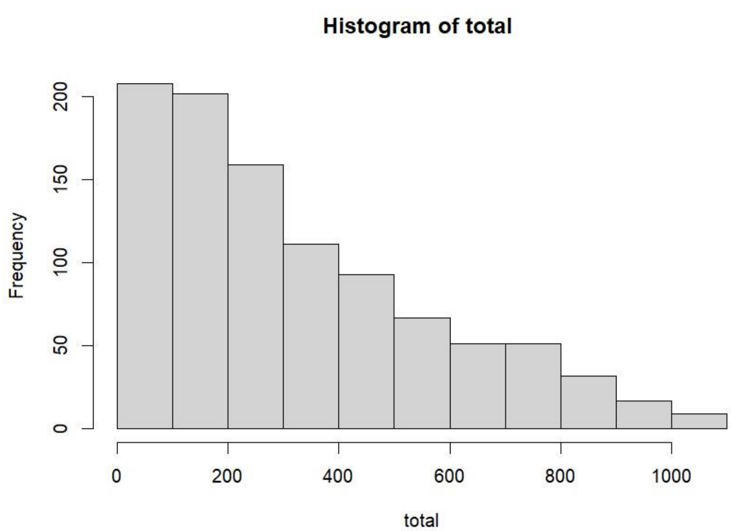
The above box plot compares the rating and product line. There are not any outliers. Seems like food and beverages has the highest mean.



The above graph shows the relationship between rating and total bill amount. We can see that points are scatted here and there so we cannot observe any pattern from this graph. The data are mostly scattered the below of the graph. Hence, we can conclude that there’s no relationship between these two variables.



The above graph represents the relationship between the tax variable and the total bill amount variable. Seems like all the points are lies on the line. So relationship between these two variables is linear.



The histogram of total shows positively skewed. This shows a decreasing frequency of totals. The customers who spend more than 500 Is considerably low. The most of customers in this supermarket chain spend a lower amount in their purchases.

5.0 Conclusion of the analysis

We drew Simple Random, Stratified and two stage cluster sample from the census data set about a supermarket sales info and estimated population parameters (mean, total and proportions to suitable variables of our choosing) using those samples. We took 2 samples per each design, and it can be observed that both samples were mostly similar. The estimates given by each sample from the respective sampling design had lower standard errors including ratio estimate (Ratio estimate we considered had and perfect correlation since the tax was calculated 5% of the total bill). Therefore, it would have been more convenient to consider a sample of SRS, Stratified or 2 stage cluster rather than taking a census to study about the population.

Comparing the results of the sample analysis results from the 3 sampling methods shows that Stratified sampling has either a roughly similar or a lower standard error than SRS. Additionally, the SE of two stage cluster sampling was considerably lower than of both SRS and stratified sampling. Therefore, it would have been more cost effective and time saving to use a two-stage cluster sample to analyze population parameters.

6.0 R code

Simple random sampling

```
#install.packages("survey")

#install.packages("sampler")

#install.packages("sampling")


#import libraries

library("survey")

library("sampler")

library("sampling")


#set working directory
```

```
setwd("E:/R stat")
mydata<-read.csv("supermarket_sales.csv")

#calculate the necessary sample size
n=rsampcalc(nrow(mydata),e=3,ci=95)
n
#Calculate the size of each stratum
sam_sz=ssampcalc(df=mydata,n=n,strata = Product.line)
sam_sz
#drawing sample 1 ------------------------------------------------------
set.seed(15548)
str_sam1=ssamp(df=mydata,n=n,strata = Product.line)
data1=data.frame(str_sam1)
data3=data.frame(sam_sz[,1],sam_sz[,3])

#include the weights
strat_sample1=merge(data3,data1,by="Product.line")
#View(strat_sample)

#estimating in stratified sample 1 .......................
attach(strat_sample1)
strat1_des=svydesign(id=~1,strata=~Product.line,weights = ~wt,data=strat_sample1)

#sample totals
#total
svytotal(~Total,strat1_des,deff=TRUE)
#total gross income
svytotal(~gross_income,strat1_des,deff=TRUE)
#total cost
```

```
svytotal(~cogs,strat1_des,deff=TRUE)

#total tax

svytotal(~Tax.5.,strat1_des,deff=TRUE)


#sample means

#total

svymean(~Total,strat1_des)

#rating

svymean(~Rating,strat1_des)

#gross income

svymean(~gross_income,strat1_des)


#sample Proportion

#customer type

svymean(~C_type,strat1_des)

#pay method

svymean(~Payment,strat1_des)

#Gender

svymean(~Gender,strat1_des)

#branch

svymean(~Branch,strat1_des)


#>>>>>>>>>>>sub groups

#mean rating by strata

svyby(~Rating,by=~Product.line,strat1_des,svymean)

#mean total by strata

svyby(~Total,by=~Product.line,strat1_des,svymean)


#total tax by strat
```

```
svyby(~Tax.5.,by=~Product.line,strat1_des,svytotal)
#---------------------------
#ratio estimation
svyratio(~Total,~Tax.5.,design =strat1_des)

detach(strat_sample1)
#++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#drawing sample 2 in stratified sampling
set.seed(15540)
str_sam2=ssamp(df=mydata,n=n,strata = Product.line)
data1=data.frame(str_sam2)
data3=data.frame(sam_sz[,1],sam_sz[,3])

#include the weights
strat_sample2=merge(data3,data1,by="Product.line")
#View(strat_sample)

#estimating in stratified sample 1 .......................
attach(strat_sample2)
strat2_des=svydesign(id=~1,strata=~Product.line,weights = ~wt,data=strat_sample2)

#sample totals
#total
svytotal(~Total,strat2_des,deff=TRUE)
#total gross income
svytotal(~gross_income,strat2_des,deff=TRUE)
#total cost
svytotal(~cogs,strat2_des,deff=TRUE)
```

```
#total tax
svytotal(~Tax.5.,strat2_des,deff=TRUE)

#sample means
#total
svymean(~Total,strat2_des)
#rating
svymean(~Rating,strat2_des)
#gross income
svymean(~gross_income,strat2_des)

#sample Proportion
#customer type
svymean(~C_type,strat2_des)
#pay method
svymean(~Payment,strat2_des)
#Gender
svymean(~Gender,strat2_des)
#branch
svymean(~Branch,strat_des)

#>>>>>>>>>>>>sub groups
#mean rating by strata
svyby(~Rating,by=~Product.line,strat2_des,svymean)
#mean total by strata
svyby(~Total,by=~Product.line,strat2_des,svymean)

#total tax by strat
svyby(~Tax.5.,by=~Product.line,strat2_des,svytotal)
```

```
#--------------------------------------------------
#ration estimation in stratified sampling

svyratio(~Total,~Tax.5.,design =strat2_des)


detach(strat_sample2)
```

Stratified sampling

```
#install.packages("survey")
#install.packages("sampler")
#install.packages("sampling")

#import libraries
library("survey")
library("sampler")
library("sampling")

#set working directory
setwd("E:/R stat")
mydata<-read.csv("supermarket_sales.csv")

#calculate the necessary sample size
n=rsampcalc(nrow(mydata),e=3,ci=95)
n
```

```
#Calculate the size of each stratum
sam_sz=ssampcalc(df=mydata,n=n,strata = Product.line)
sam_sz
#drawing sample 1 ----------------------------------------------------
set.seed(15548)
str_sam1=ssamp(df=mydata,n=n,strata = Product.line)
data1=data.frame(str_sam1)
data3=data.frame(sam_sz[,1],sam_sz[,3])

#include the weights
strat_sample1=merge(data3,data1,by="Product.line")
#View(strat_sample)

#estimating in stratified sample 1 .......................
attach(strat_sample1)
strat1_des=svydesign(id=~1,strata=~Product.line,weights = ~wt,data=strat_sample1)

#sample totals
#total
svytotal(~Total,strat1_des,deff=TRUE)
#total gross income
svytotal(~gross_income,strat1_des,deff=TRUE)
#total cost
svytotal(~cogs,strat1_des,deff=TRUE)
#total tax
svytotal(~Tax.5.,strat1_des,deff=TRUE)

#sample means
#total
```

```
svymean(~Total,strat1_des)

#rating

svymean(~Rating,strat1_des)

#gross income

svymean(~gross_income,strat1_des)


#sample Proportion

#customer type

svymean(~C_type,strat1_des)

#pay method

svymean(~Payment,strat1_des)

#Gender

svymean(~Gender,strat1_des)

#branch

svymean(~Branch,strat1_des)


#>>>>>>>>>>>sub groups

#mean rating by strata

svyby(~Rating,by=~Product.line,strat1_des,svymean)

#mean total by strata

svyby(~Total,by=~Product.line,strat1_des,svymean)


#total tax by strat

svyby(~Tax.5.,by=~Product.line,strat1_des,svytotal)

#--------------------------

#ratio estimation

svyratio(~Total,~Tax.5.,design =strat1_des)


detach(strat_sample1)
```

```
#++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#drawing sample 2 in stratified sampling
set.seed(15540)
str_sam2=ssamp(df=mydata,n=n,strata = Product.line)
data1=data.frame(str_sam2)
data3=data.frame(sam_sz[,1],sam_sz[,3])

#include the weights
strat_sample2=merge(data3,data1,by="Product.line")
#View(strat_sample)

#estimating in stratified sample 1 .......................
attach(strat_sample2)
strat2_des=svydesign(id=~1,strata=~Product.line,weights = ~wt,data=strat_sample2)

#sample totals
#total
svytotal(~Total,strat2_des,deff=TRUE)
#total gross income
svytotal(~gross_income,strat2_des,deff=TRUE)
#total cost
svytotal(~cogs,strat2_des,deff=TRUE)
#total tax
svytotal(~Tax.5.,strat2_des,deff=TRUE)

#sample means
#total
svymean(~Total,strat2_des)
```

```
#rating
svymean(~Rating,strat2_des)
#gross income
svymean(~gross_income,strat2_des)

#sample Proportion
#customer type
svymean(~C_type,strat2_des)
#pay method
svymean(~Payment,strat2_des)
#Gender
svymean(~Gender,strat2_des)
#branch
svymean(~Branch,strat_des)

#>>>>>>>>>>>>sub groups
#mean rating by strata
svyby(~Rating,by=~Product.line,strat2_des,svymean)
#mean total by strata
svyby(~Total,by=~Product.line,strat2_des,svymean)

#total tax by strat
svyby(~Tax.5.,by=~Product.line,strat2_des,svytotal)

#--------------------------------------------------
#ration estimation in stratified sampling

svyratio(~Total,~Tax.5.,design =strat2_des)
```

```
detach(strat_sample2)
```

Cluster sampling

```
#install.packages("survey")
#install.packages("sampler")
#install.packages("sampling")

#importing the libraries
library("survey")
library("sampler")
library("sampling")

#set working directry
setwd("E:/R stat")
mydata<-read.csv("supermarket_sales.csv")

#------------------------------------------------------------------
#drawing sample 1 in cluster
#first select few clusters randomly
set.seed(15550)
```

```
cl = cluster(mydata,clustername = "City",size = 2,method = "srswor",description = T)
clus_st1 = getdata(mydata,cl)
t1 = table(clus_st1$City)
t1
#View(clus_st2)



#getting srs from selected clusters
cities = names(t1)
clus_st2 = data.frame()

for (i in cities) {
  srs_size=rsampcalc(nrow(clus_st1[clus_st1$City==i,]),e=3,ci=95) #sample size for SRS
  stg2 = clus_st1[clus_st1$City==i,][sample(1:t1[i],srs_size,replace=FALSE),]

  clus_st2 = rbind(clus_st2,stg2)
}
#head(clus_st2)
#View(clus_st2) # = 2-stage cluster sample
t2=table(clus_st2$City)
t2
t2/t1
##weights calculating
N=3
n=2
weighted_val = c()
for (i in 1:sum(t2)) {
  city = clus_st2[i,"City"]
  weighted_val[i] = (N*t1[city])/(n*t2[city])
```

```
}
clus_st2 = cbind(clus_st2,weighted_val)
#head(clus_st2)
#tail(stage_2_2nd)
#View(stage_2_2nd)

#esitmating in two stage cluster sample 1-----------------------------------
attach(clus_st2)
clus_des = svydesign(id=~City, weights = ~weighted_val, data = clus_st2)
clus_des

#sample_means
#mean total
svymean(~Total,clus_des,deff=TRUE)
#mean rating
svymean(~Rating,clus_des,deff=TRUE)
#gross income
svymean(~gross_income,clus_des,deff=TRUE)

#sample_totals
#total tax
svytotal(~Tax.5.,clus_des)
#total cost
svytotal(~cogs,clus_des)
#total gross income
svytotal(~gross_income,clus_des)

#sample_proportions
#gender
```

```
svymean(~Gender,clus_des)


#sample Proportion
#customer type
table(C_type)/length(C_type)
#pay method
table(Payment)/length(Payment)
#Gender
table(Gender)/length(Gender)
#branch
table(Branch)/length(Branch)


#sample Proportion
#customer type
svymean(~C_type,clus_des)
#pay method
svymean(~Payment,clus_des)
#Gender
svymean(~Gender,clus_des)
#branch
svymean(~Branch,clus_des)
#ratio estimation.........................................................


m_i = 300
t_i_hat_2 = c()
for (i in 1:600) {
  sec = stage_2_2nd[i,"Sector"]
  t_i_hat_2[i] = (t1[sec]*stage_2_2nd$Income.in.thousand[i])/(m_i)
}
```

```
stage_2_2nd = cbind(stage_2_2nd,t_i_hat_2)
head(stage_2_2nd)


t1 = table(stage_1_2nd$Sector)
#Ratio estimation for income_sample 1
yr_bar_hat = sum(stage_2_2nd$t_i_hat_2)/sum(t1)
yr_bar_hat

detach(clus_st2)


#++++++++++++++++++,,,,,,,,,,,,,,,,,,,,,,,,,,++++++++++++++++++++++++++++++++++++++++

#drawing sample 2 in  cluster
#first select few clusters randomly
set.seed(15561)
cl2 = cluster(mydata,clustername = "City",size = 2,method = "srswor",description = T)
clus2_st1 = getdata(mydata,cl2)
t2 = table(clus2_st1$City)
t2
#View(clus_st2)


#getting srs from selected clusters
cities = names(t2)
clus2_st2 = data.frame()

for (i in cities) {
```

```
  srs_size2=rsampcalc(nrow(clus2_st1[clus2_st1$City==i,]),e=3,ci=95) #sample size for SRS
  stg2 = clus2_st1[clus2_st1$City==i,][sample(1:t2[i],srs_size2,replace=FALSE),]

  clus2_st2 = rbind(clus2_st2,stg2)
}
#head(clus2_st2)
#View(clus2_st2) # = 2-stage cluster sample
tb=table(clus2_st2$City)
tb


##weights calculating
N=3
n=2
weighted_val2 = c()
for (i in 1:sum(tb)) {
  city = clus2_st2[i,"City"]
  weighted_val2[i] = (N*t2[city])/(n*tb[city])
}
clus2_st2 = cbind(clus2_st2,weighted_val2)
#head(clus2_st2)
#tail(clus2_st2)
#View(clus2_st2)

#esitmating in two stage cluster sample 1----------------------------------
attach(clus2_st2)
clus2_des = svydesign(id=~City, weights = ~weighted_val2, data = clus2_st2)
clus2_des
```

```
#sample_means
#mean total
svymean(~Total,clus2_des,deff=TRUE)
#mean rating
svymean(~Rating,clus2_des,deff=TRUE)
#gross income
svymean(~gross_income,clus2_des,deff=TRUE)

#sample_totals
#total tax
svytotal(~Tax.5.,clus2_des)
#total cost
svytotal(~cogs,clus2_des)
#total gross income
svytotal(~gross_income,clus2_des)

#sample_proportions
#sample Proportion
#customer type
table(C_type)/length(C_type)
#pay method
table(Payment)/length(Payment)
#Gender
table(Gender)/length(Gender)
#branch
table(Branch)/length(Branch)

#sample Proportion
#customer type
```

```
svymean(~C_type,clus2_des)

#pay method

svymean(~Payment,clus2_des)

#Gender

svymean(~Gender,clus2_des)

#branch

svymean(~Branch,clus2_des)



detach(clus2_st2)
```

Graphical analysis code

```
# Set the working directory to where your CSV file is located

setwd("D:/Campus/Level 3/Sem 1/IS 3001/Grp.Proj/Project")

# Read the CSV file

mydata <- read.csv("supermarket_sales.csv")

data1=data.frame(mydata)

#View(Data)

t1=table(mydata$Product.line)

prod=names(t1)

prod

table(mydata$City)


#barchart

barplot(table(mydata$Branch), names.arg=c("A","B","C"), col="blue",main="Number of branches
",ylab="count")
```

```
barplot(table(mydata$Payment),main="Payment method use",
        names.arg =c("Ewallet","Cash","Credit card"),
        ylab = "Proportion" )

#boxplot
pop_des<-svydes()
boxplot(Rating~Product.line,data1,all.outliers = TRUE)

#scatterplot
plot(mydata$Total,mydata$rating,
     main="Scatterplot of ratings and total bill amount",
     xlab="total amount",ylab="ratings")
plot(mydata$Total,mydata$Tax.5.,
     main="Scatterplot of tax and total bill amount",
     xlab="total amount",ylab="tax")

plot(mydata$Quantity,mydata$Total,
     main="Scatterplot of tax and total bill amount",
     xlab="Quantity",ylab="Total")

#histogram
hist(x = mydata$Total,prob=F,main="Histogram of total ",xlab="
total")
```