Assignment 2

Dixitha Kasturi
dkasturi@syr.edu

The supermarket dataset has 1000 records of purchases with 18 different attributes for each purchase. We will be exploring some of these attributes in detail to understand and get insights from the data that could be further used to understand any underlying conditions or situations that are resulting in the numbers that are reflected.

EXPLORATORY DATA ANALYSIS:

Overall it is a clean dataset with not a lot of transformations to do. We don't necessarily see the need to transform any attribute. Customer type, Gender, Product line, Payment, Branch are all character type, but we don't need to convert them into factors despite them having repetitive values, is because factors in R are nothing but ordinal data. And we don't need order/hierarchy for any of the above-mentioned attributes. So i am not performing any transformation.

```
spec_tbl_df [1,000 x 17] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Invoice ID             : chr [1:1000] "750-67-8428" "226-31-3081" "631-41-3108" "123-19-
1176" ...
 $ Branch                 : chr [1:1000] "A" "C" "A" "A" ...
 $ City                   : chr [1:1000] "Yangon" "Naypyitaw" "Yangon" "Yangon" ...
 $ Customer type          : chr [1:1000] "Member" "Normal" "Normal" "Member" ...
 $ Gender                 : chr [1:1000] "Female" "Female" "Male" "Male" ...
 $ Product line           : chr [1:1000] "Health and beauty" "Electronic accessories" "Home
and lifestyle" "Health and beauty" ...
 $ Unit price             : num [1:1000] 74.7 15.3 46.3 58.2 86.3 ...
 $ Quantity               : num [1:1000] 7 5 7 8 7 7 6 10 2 3 ...
 $ Tax 5%                 : num [1:1000] 26.14 3.82 16.22 23.29 30.21 ...
 $ Total                  : num [1:1000] 549 80.2 340.5 489 634.4 ...
 $ Date                   : chr [1:1000] "1/5/2019" "3/8/2019" "3/3/2019" "1/27/2019" ...
 $ Time                   : 'hms' num [1:1000] 13:08:00 10:29:00 13:23:00 20:33:00 ...
  ..- attr(*, "units")= chr "secs"
 $ Payment                : chr [1:1000] "Ewallet" "Cash" "Credit card" "Ewallet" ...
 $ cogs                   : num [1:1000] 522.8 76.4 324.3 465.8 604.2 ...
 $ gross margin percentage: num [1:1000] 4.76 4.76 4.76 4.76 4.76 ...
 $ gross income           : num [1:1000] 26.14 3.82 16.22 23.29 30.21 ...
 $ Rating                 : num [1:1000] 9.1 9.6 7.4 8.4 5.3 4.1 5.8 8 7.2 5.9 ...
```

SUMMARY:

```
  Invoice ID           Branch             City           Customer type
 Length:1000        Length:1000        Length:1000        Length:1000
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



    Gender          Product line         Unit price       Quantity        Tax 5%
 Length:1000        Length:1000        Min.   :10.08   Min.   : 1.00   Min.   : 0.5085
 Class :character   Class :character   1st Qu.:32.88   1st Qu.: 3.00   1st Qu.: 5.9249
 Mode  :character   Mode  :character   Median :55.23   Median : 5.00   Median :12.0880
                                       Mean   :55.67   Mean   : 5.51   Mean   :15.3794
                                       3rd Qu.:77.94   3rd Qu.: 8.00   3rd Qu.:22.4453
                                       Max.   :99.96   Max.   :10.00   Max.   :49.6500
     Total             Date              Time           Payment
 Min.   :  10.68   Length:1000        Length:1000        Length:1000
 1st Qu.: 124.42   Class :character   Class1:hms         Class :character
 Median : 253.85   Mode  :character   Class2:difftime    Mode  :character
 Mean   : 322.97                      Mode  :numeric
 3rd Qu.: 471.35
 Max.   :1042.65
      cogs        gross margin percentage  gross income         Rating
 Min.   : 10.17   Min.   :4.762           Min.   : 0.5085   Min.   : 4.000
 1st Qu.:118.50   1st Qu.:4.762           1st Qu.: 5.9249   1st Qu.: 5.500
 Median :241.76   Median :4.762           Median :12.0880   Median : 7.000
 Mean   :307.59   Mean   :4.762           Mean   :15.3794   Mean   : 6.973
 3rd Qu.:448.90   3rd Qu.:4.762           3rd Qu.:22.4453   3rd Qu.: 8.500
 Max.   :993.00   Max.   :4.762           Max.   :49.6500   Max.   :10.000
```

DESCRIBE :

Describe gives us complete picture of the attributes from the missing values, distinct values , descriptive statistics and quantile values. There are no missing values in any of the attributes. Almost equal number of orders from males and females, Members and normal customers. From the descriptive statistics especially for the numeric attributes, we see that the values are the same for Tax and Grossincome. An interesting observation is that the lowest rating was 4 on a scale of 1-10. For time(from 10am to 9pm) since the values are continuous, by discretizing it we can give a broader picture of our analysis related to time.

```
market

 17  Variables      1000  Observations
---------------------------------------------------------------------------
Invoice ID
       n  missing distinct
    1000        0     1000

lowest : 101-17-6199 101-81-4070 102-06-2002 102-77-2261 105-10-6182
highest: 894-41-5205 895-03-6665 895-66-0685 896-34-0956 898-04-2717
---------------------------------------------------------------------------
Branch
       n  missing distinct
    1000        0        3

Value          A     B     C
Frequency    340   332   328
Proportion 0.340 0.332 0.328
---------------------------------------------------------------------------
City
       n  missing distinct
    1000        0        3

Value       Mandalay Naypyitaw    Yangon
Frequency        332       328       340
Proportion     0.332     0.328     0.340
---------------------------------------------------------------------------
Customer type
       n  missing distinct
    1000        0        2

Value      Member Normal
Frequency     501    499
Proportion  0.501  0.499
---------------------------------------------------------------------------
Gender
       n  missing distinct
    1000        0        2

Value      Female   Male
Frequency     501    499
Proportion  0.501  0.499
---------------------------------------------------------------------------
Product line
       n  missing distinct
    1000        0        6

lowest : Electronic accessories Fashion accessories    Food and beverages     Health and
beauty      Home and lifestyle
highest: Fashion accessories    Food and beverages     Health and beauty      Home and
lifestyle     Sports and travel

Value      Electronic accessories    Fashion accessories     Food and beverages
Frequency                     170                     178                    174
Proportion                  0.170                   0.178                  0.174
---------------------------------------------------------------------------
```

```
Product line
       n  missing distinct
    1000        0        6

lowest : Electronic accessories Fashion accessories    Food and beverages     Health and
beauty      Home and lifestyle
highest: Fashion accessories    Food and beverages     Health and beauty      Home and
lifestyle     Sports and travel

Value      Electronic accessories    Fashion accessories     Food and beverages
Frequency                     170                     178                     174
Proportion                  0.170                   0.178                   0.174

Value           Health and beauty     Home and lifestyle      Sports and travel
Frequency                     152                     160                     166
Proportion                  0.152                   0.160                   0.166
-------------------------------------------------------------------------------------------
Unit price
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
    1000        0      943        1    55.67     30.6    15.28    19.31    32.88    55.23
     .75      .90      .95
   77.94    93.12    97.22

lowest : 10.08 10.13 10.16 10.17 10.18, highest: 99.82 99.83 99.89 99.92 99.96
```

```
Quantity
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
    1000        0       10     0.99     5.51     3.36        1        1        3        5
     .75      .90      .95
       8       10       10

lowest :  1  2  3  4  5, highest:  6  7  8  9 10

Value          1     2     3     4     5     6     7     8     9    10
Frequency    112    91    90   109   102    98   102    85    92   119
Proportion 0.112 0.091 0.090 0.109 0.102 0.098 0.102 0.085 0.092 0.119
-------------------------------------------------------------------------------------------
Tax 5%
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
    1000        0      990        1    15.38    12.89    1.956    3.243    5.925   12.088
     .75      .90      .95
  22.445   34.234   39.166

lowest :  0.5085  0.6045  0.6270  0.6390  0.6990, highest: 48.6900 48.7500 49.2600 49.4900
49.6500
-------------------------------------------------------------------------------------------
Total
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
    1000        0      990        1      323    270.7    41.07    68.10   124.42   253.85
     .75      .90      .95
  471.35   718.91   822.50

lowest :   10.6785   12.6945   13.1670   13.4190   14.6790
highest: 1022.4900 1023.7500 1034.4600 1039.2900 1042.6500
```

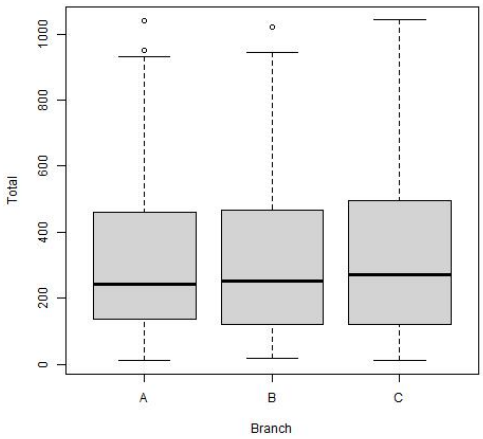
```
Date
       n  missing distinct
    1000        0       89

lowest : 1/1/2019  1/10/2019 1/11/2019 1/12/2019 1/13/2019
highest: 3/5/2019  3/6/2019  3/7/2019  3/8/2019  3/9/2019
-------------------------------------------------------------------------------------------
Time [secs]
       n  missing distinct
    1000        0      506

lowest : 10:00:00 10:01:00 10:02:00 10:03:00 10:04:00, highest: 20:52:00 20:54:00 20:55:00
20:57:00 20:59:00
-------------------------------------------------------------------------------------------
Payment
       n  missing distinct
    1000        0        3

Value             Cash Credit card     Ewallet
Frequency          344         311         345
Proportion       0.344       0.311       0.345
-------------------------------------------------------------------------------------------
cogs
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
    1000        0      990        1    307.6    257.8    39.11    64.86   118.50   241.76
     .75      .90      .95
  448.91   684.68   783.33

lowest :  10.17  12.09  12.54  12.78  13.98, highest: 973.80 975.00 985.20 989.80 993.00
```

```
gross margin percentage
       n  missing distinct     Info     Mean      Gmd
    1000        0        1        0    4.762        0

Value      4.761905
Frequency      1000
Proportion        1
-------------------------------------------------------------------------------------------
gross income
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
    1000        0      990        1    15.38    12.89    1.956    3.243    5.925   12.088
     .75      .90      .95
  22.445   34.234   39.166

lowest :  0.5085  0.6045  0.6270  0.6390  0.6990, highest: 48.6900 48.7500 49.2600 49.4900
49.6500
-------------------------------------------------------------------------------------------
Rating
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
    1000        0       61        1    6.973    1.985    4.295    4.500    5.500    7.000
     .75      .90      .95
   8.500    9.400    9.700

lowest :  4.0  4.1  4.2  4.3  4.4, highest:  9.6  9.7  9.8  9.9 10.0
-------------------------------------------------------------------------------------------
```

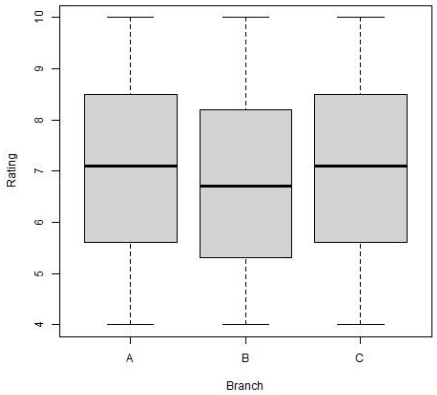
OUTLIERS :

Just a few outliers in the total cost from branch A and B. These values dont need to be excluded from the dataset as they are less in number and they can be justified as high purchases.

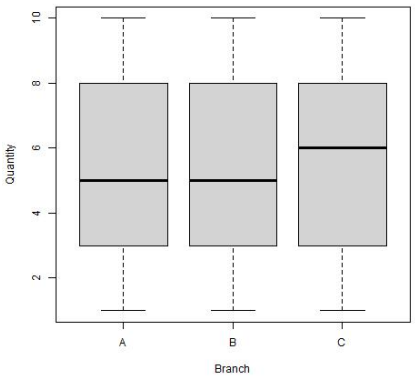
a) Sales by branch



b) Rating by branch



c)Quantity sold by branch

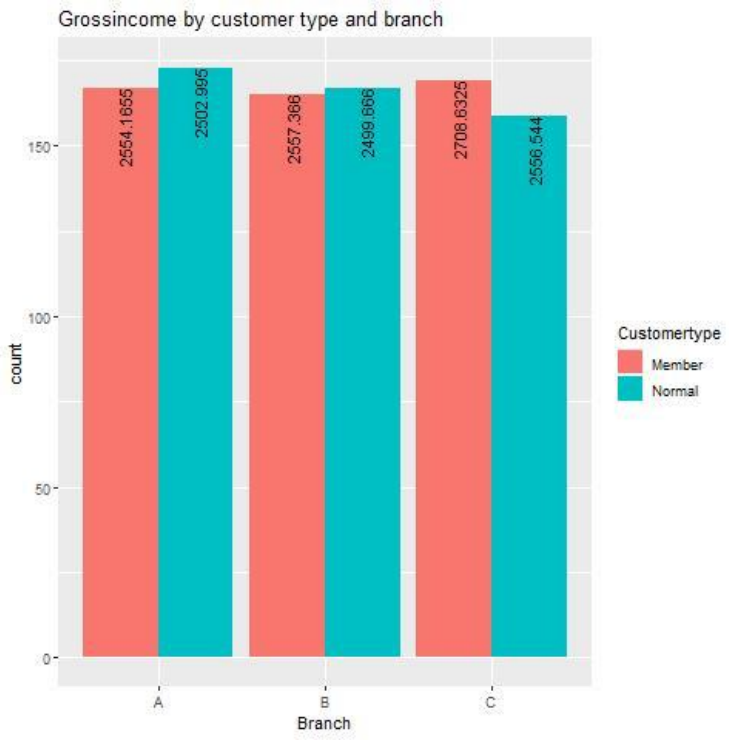


CLEANING:

```
#Cleaning column names for easy use

colnames(market) <- c('InvoiceID','Branch','City','Customertype','Gender',
                      'Productline','Unitprice','Quantity','Tax','Total',
                      'Date','Time','Payment','cogs','grossmarginpercentage',
                      'grossincome','Rating')


# Discretizing time levels
market$time_levels <- cut(strptime(market$Time, format = "%H:%M:%S"),
                          breaks = strptime(c( "10:00:00","12:00:00","16:00:00",
                                               "19:00:00","21:00:00")
                                            , format = "%H:%M:%S"),
   labels = c("Morning", "Afternoon","Evening","Night"))


# Getting data by branches

branchA <- market[which(market$Branch == 'A'),]
branchB <- market[which(market$Branch == 'B'),]
branchC <- market[which(market$Branch == 'C'),]
```

ANALYSIS

Grossincome by customertype and branch - All three branches had similar gross income generated by customers who were members and normal type. Gross income and tax summed upto the same, which was first observed in descriptive statistics(2nd figure).

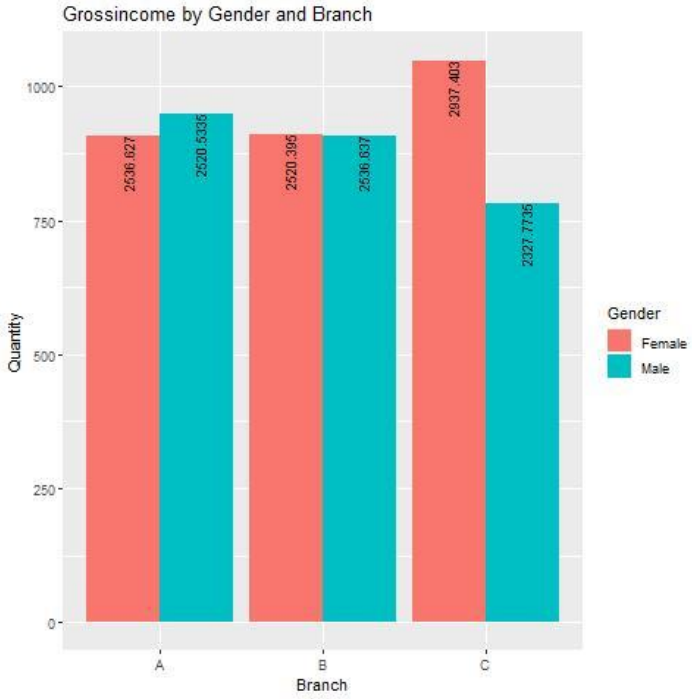


A tibble: 3 x 4

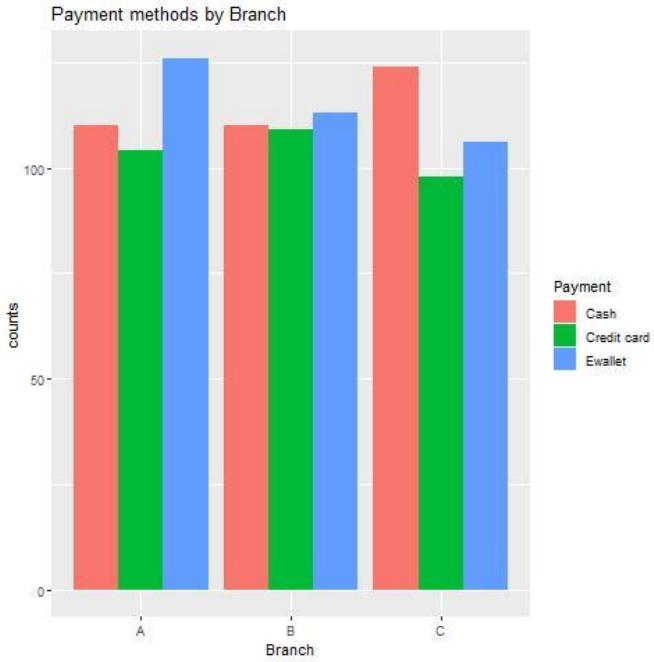
Branch <chr>	grossincome <dbl>	Quantity <dbl>	Tax <dbl>
A	5057.160	1859	5057.160
B	5057.032	1820	5057.032
C	5265.176	1831	5265.176

Grossincome by gender type and branch

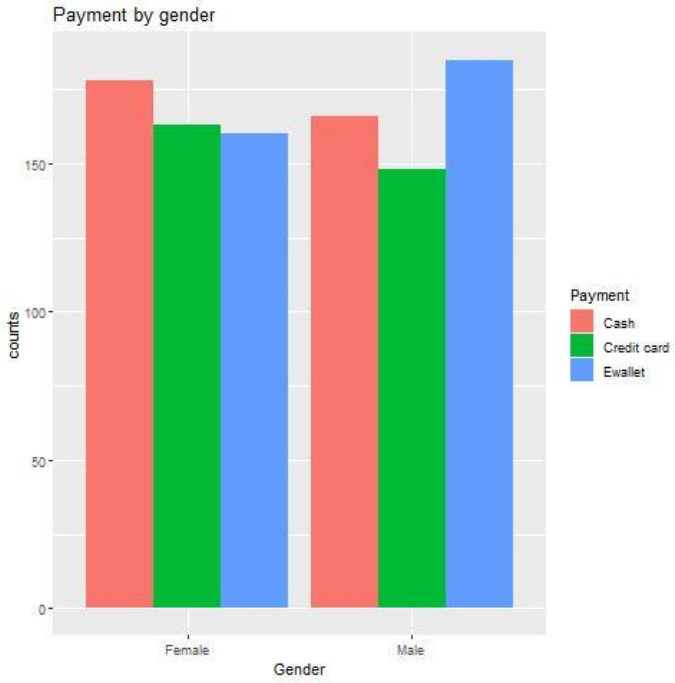
Branch A : Males generated high Gross income($2520.53).
Branch B : Males and Females generated equal gross income
Branch C : Females generated high gross income($2937.4).



Payment methods used by branch - At Branch A, Ewallet high Ewallet payments were made, while at Branch B all 3 payment methods were equally used. On the contrary Branch C had high cash payments, this does in a way indicate that change should be checked often.



Payment methods used by Gender: Females mostly preferred Cash payments while males preferred Ewallet

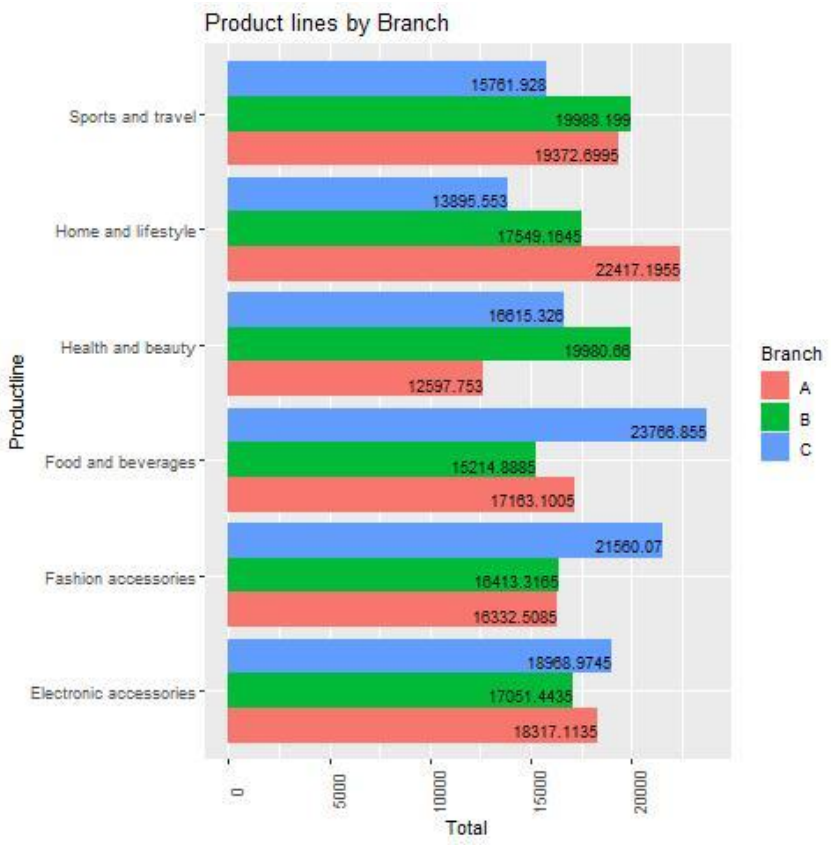


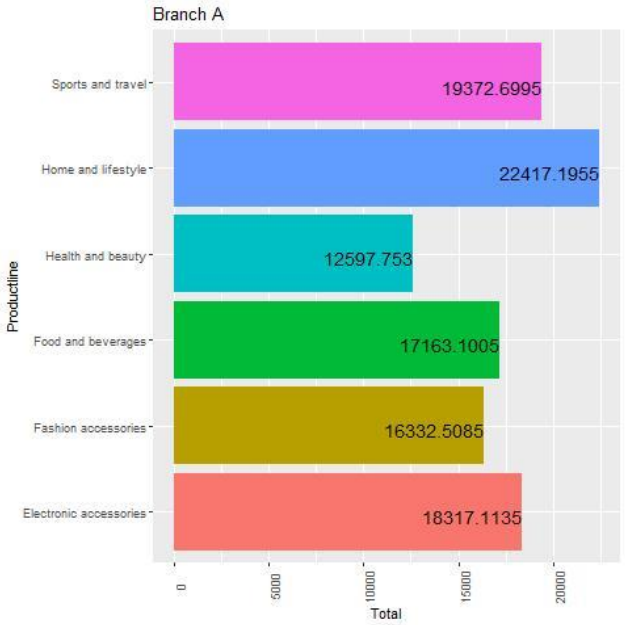
By Branch findings:

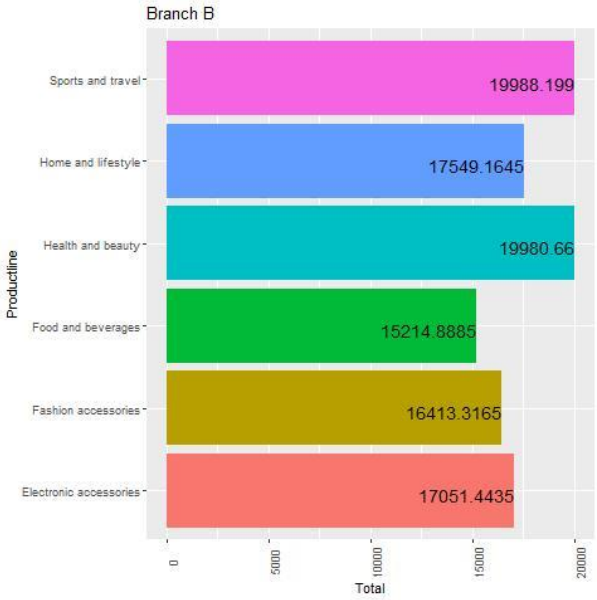
Branch A generated high total amounts in home and lifestyle($22417.19)

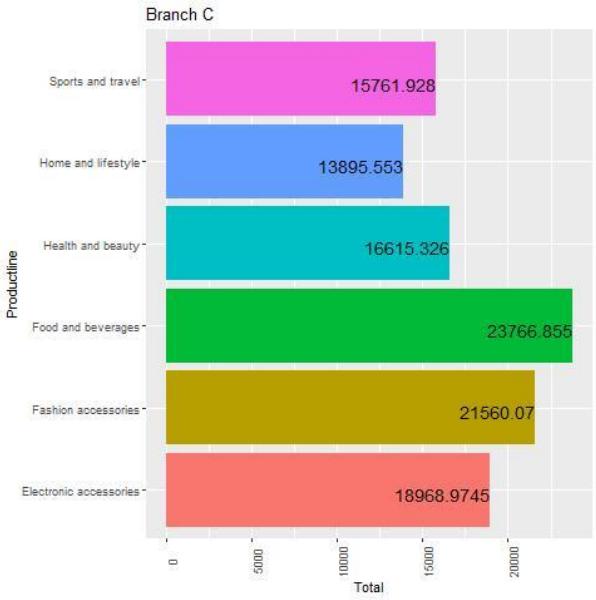
Branch B generated high total amounts in Sports and travel($19988.19).Health and beauty($19980.66)

Branch C generated high total amounts in Food and Beverages($23766.85) , Fashion accessories ($21560.0), Electronic accessories($18968.97)









By Gender findings -

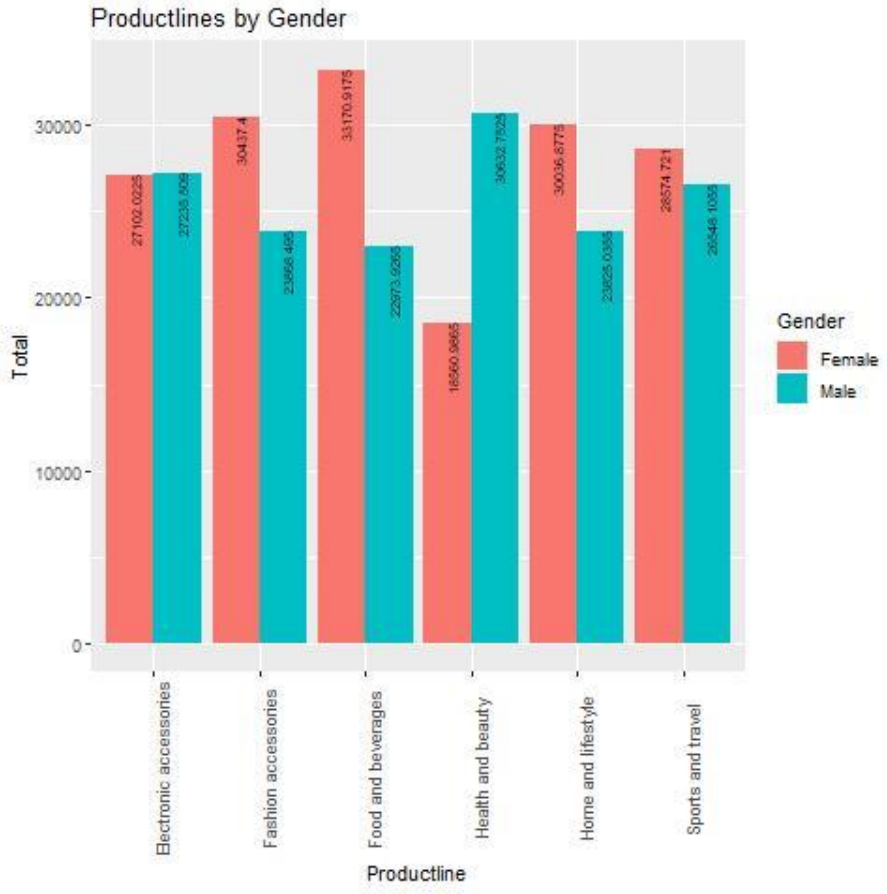
Branch A : Females purchased more electronic accessories, fashion accessories, home and lifestyle products while males purchased more in Food and Beverages, sports and travel and surprisingly health and beauty

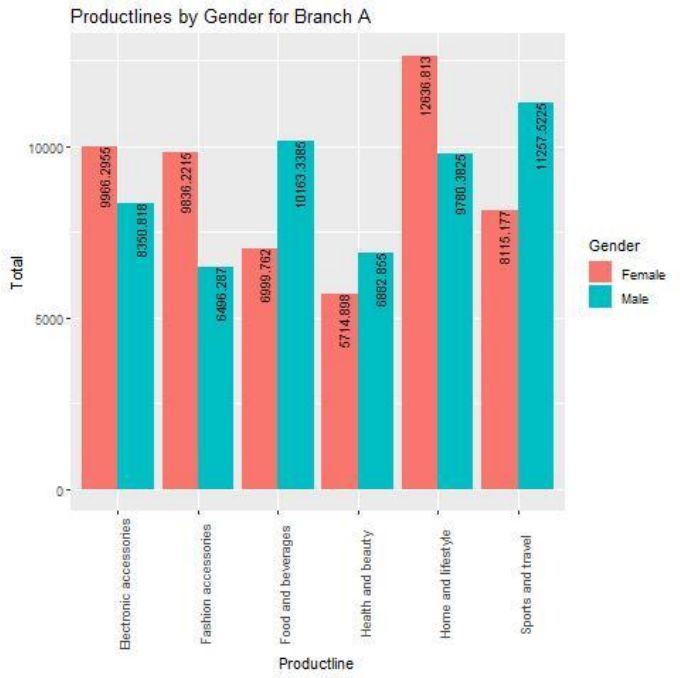
Branch B : Females purchased more Food and Beverages, fashion accessories, home and lifestyle products while males purchased more in electronic accessories, sports and travel and surprisingly very high in health and beauty

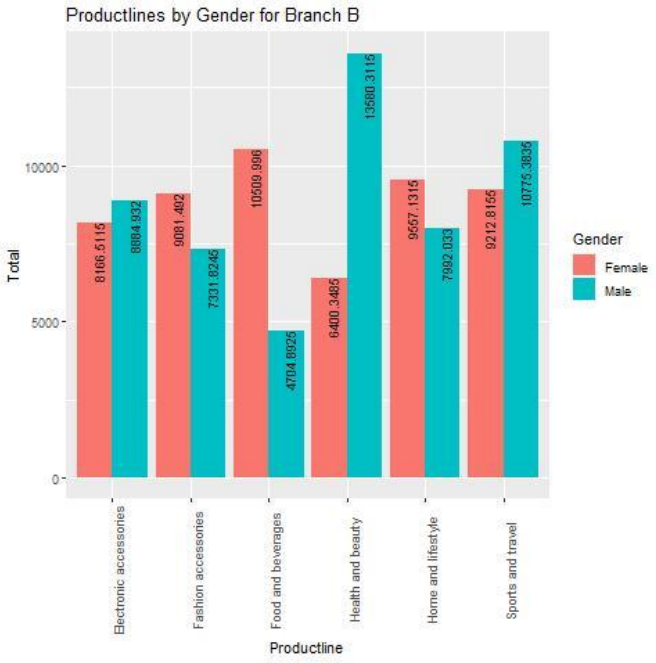
Branch C : Females purchased more Food and Beverages, fashion accessories, home and lifestyle, sports and travel products while males purchased more in electronic accessories, and surprisingly very high in health and beauty

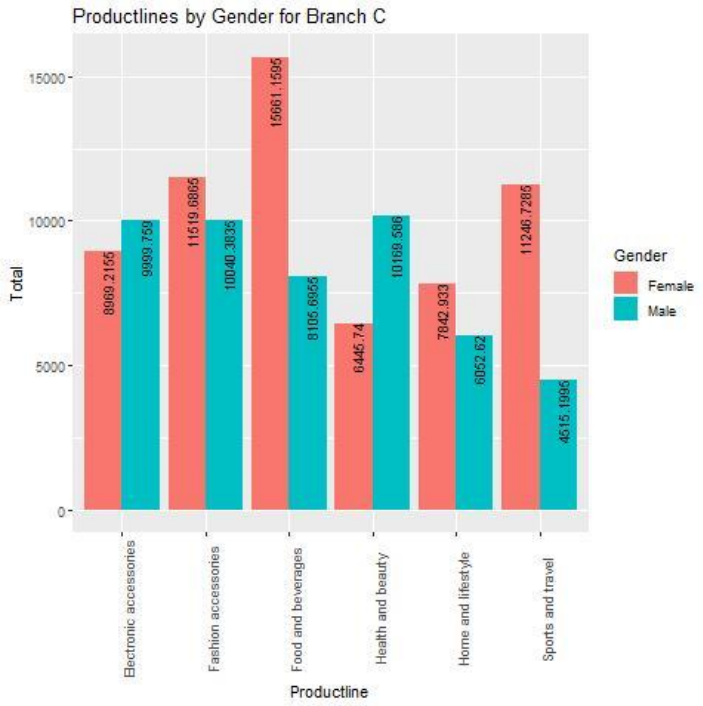
Overall females purchased more products from Food and Beverages, fashion accessories, home and lifestyle, sports and travel products

Overall males purchased more in health and beauty and a little more in electronic goods.


Productlines by Gender
Total
30000
20000
10000
0
27102.0225
27235.909
30437.4
23966.495
33170.9175
22973.9205
18560.9665
30932.7523
30036.6775
23825.0385
28574.721
26543.1055
Gender
Female
Male
Electronic accessories
Fashion accessories
Food and beverages
Health and beauty
Home and lifestyle
Sports and travel
Productline

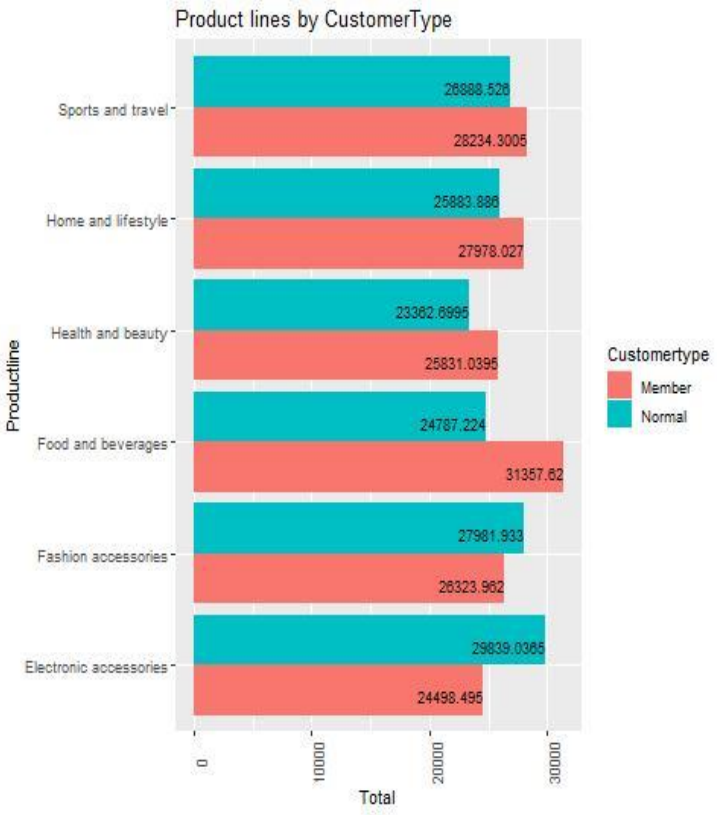

Productlines by Gender for Branch A
Total
10000
5000
0
9966.2955
8350.818
9836.2215
6496.287
6999.762
10163.3385
5714.898
6862.865
12636.813
9780.3825
8115.177
11257.5225
Gender
Female
Male
Electronic accessories
Fashion accessories
Food and beverages
Health and beauty
Home and lifestyle
Sports and travel
Productline


Productlines by Gender for Branch B
Total
10000
5000
0
8166.5115
8884.932
9081.492
7331.8245
10509.996
4704.8925
6400.3485
13580.3115
9567.1315
7992.033
9212.8155
10775.3835
Gender
Female
Male
Electronic accessories
Fashion accessories
Food and beverages
Health and beauty
Home and lifestyle
Sports and travel
Productline



Productlines by Customertype:

Overall, customers who are existing members in contributed more to Sports and travel, home and lifestyle, health and beauty, food and beverages. Normal customers, who were not members, had high sales in Fashion accessories and electronic accessories.

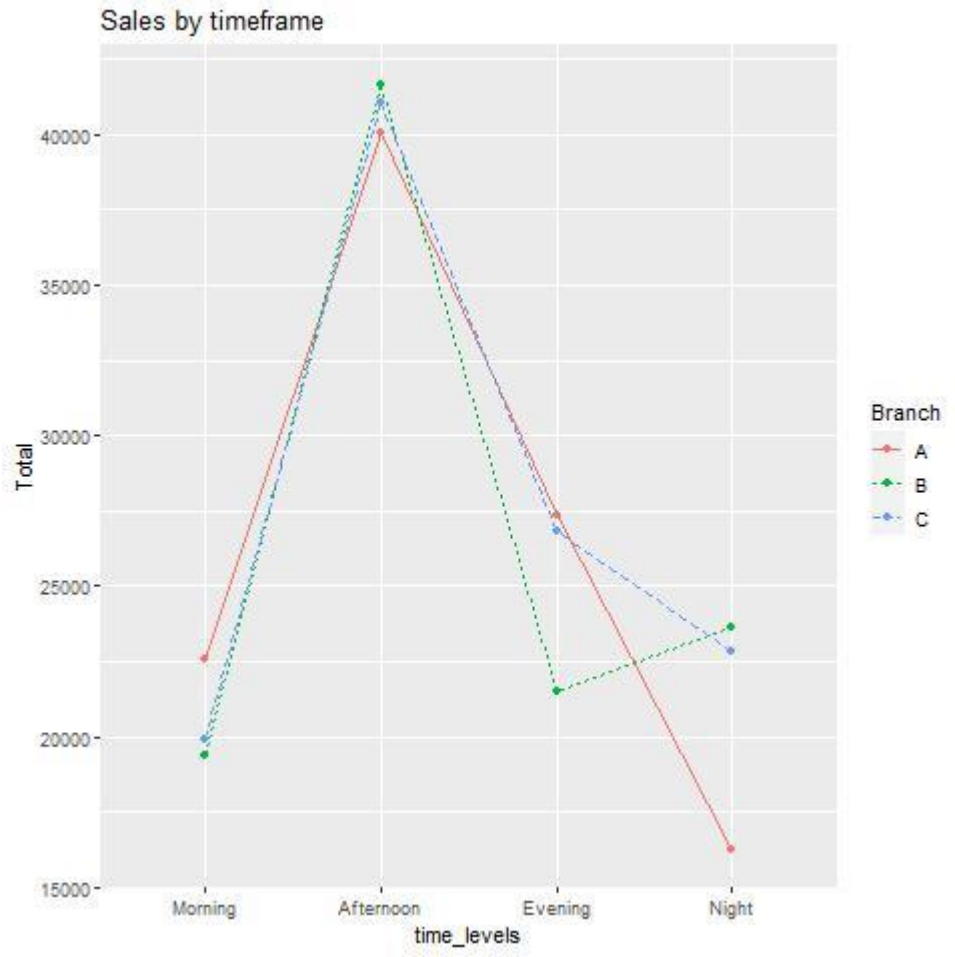


Customertype <chr>	Productline <chr>	counts <int>	Total <dbl>
Member	Electronic accessories	78	24498.49
Member	Fashion accessories	86	26323.96
Member	Food and beverages	94	31357.62
Member	Health and beauty	73	25831.04
Member	Home and lifestyle	83	27978.03
Member	Sports and travel	87	28234.30
Normal	Electronic accessories	92	29839.04
Normal	Fashion accessories	92	27981.93
Normal	Food and beverages	80	24787.22
Normal	Health and beauty	79	23362.70

Customertype <chr>	Productline <chr>	counts <int>	Total <dbl>
Normal	Home and lifestyle	77	25883.89
Normal	Sports and travel	79	26888.53

Timeperiod of Sales -

Discretized Timings of orders into 4 groups : 10 am to 12pm as morning, 12pm to 4pm as afternoon, 4pm to 7pm as evening and 7pm to 9pm as night In all 3 branches, the total sales and gross income were high during afternoon period and sales gradually increased from 9am to 12pm
For Branch A : Sales continuously dipped after 4pm
For Branch B : Sales dipped from 4pm to 7pm but increased after 7pm
For Branch C : Sales decreased after 4pm but it was not a steep drop.



Branch <chr>	time_levels <fctr>	counts <int>	grossincome <dbl>	Total <dbl>
A	Morning	73	1074.2050	22558.31
A	Afternoon	126	1907.3780	40054.94
A	Evening	92	1302.8805	27360.49
A	Night	49	772.6970	16226.64
B	Morning	59	921.3170	19347.66
B	Afternoon	125	1984.9310	41683.55
B	Evening	72	1024.7570	21519.90
B	Night	76	1126.0270	23646.57
C	Morning	59	947.2785	19892.85
C	Afternoon	126	1955.1680	41058.53

Branch <chr>	time_levels <fctr>	counts <int>	grossincome <dbl>	Total <dbl>
C	Evening	80	1277.2140	26821.49
C	Night	63	1085.5160	22795.84

Rating analysis -

by store, by category, by gender, by customertype

The average rating given for all 3 branches were more or less the same but branch C had the highest at 7.07

Customers who were members surprisingly had low average rating at 6.94 compared to normal customers with average of 7

Male and Female customers had more or less equal average rating overall but for individual branches, Males had high average rating for branch A while females had high average rating for branch B,C

Females gave higher rating for health and beauty, fashion and accessories, food and beverages and categories whereas males gave higher ratings in electronic accessories, home and lifestyle, sports and travel

Branch <chr>	counts <int>	Averagerating <dbl>
A	340	7.027059
B	332	6.818072
C	328	7.072866

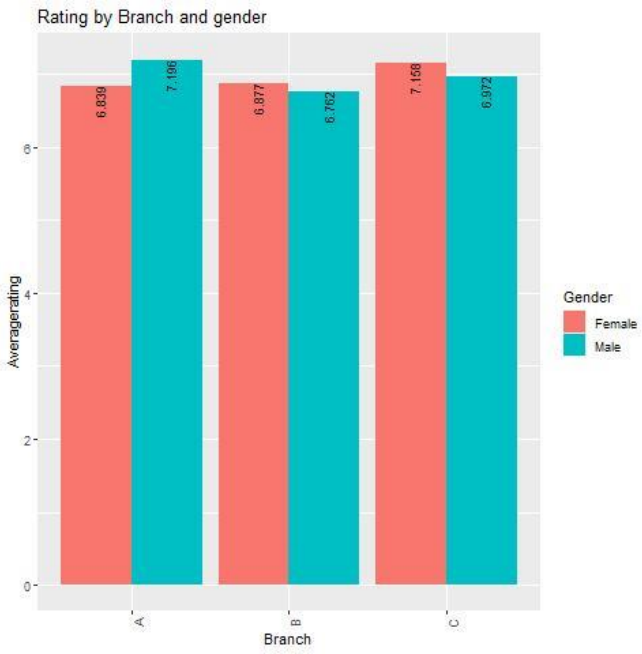
Customertype <chr>	counts <int>	Averagerating <dbl>
Member	501	6.940319
Normal	499	7.005210

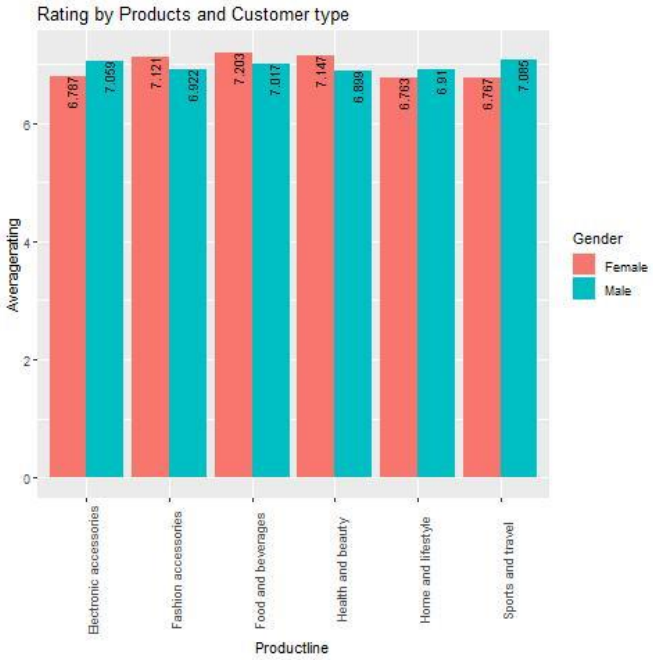
Gender <chr>	counts <int>	Averagerating <dbl>
Female	501	6.964471
Male	499	6.980962

Branch <chr>	Gender <chr>	counts <int>	Averagerating <dbl>
A	Female	161	6.839
A	Male	179	7.196
B	Female	162	6.877
B	Male	170	6.762
C	Female	178	7.158
C	Male	150	6.972

Gender <chr>	Productline <chr>	counts <int>	Averagerating <dbl>
Female	Electronic accessories	84	6.787
Female	Fashion accessories	96	7.121
Female	Food and beverages	90	7.203
Female	Health and beauty	64	7.147
Female	Home and lifestyle	79	6.763
Female	Sports and travel	88	6.767
Male	Electronic accessories	86	7.059
Male	Fashion accessories	82	6.922
Male	Food and beverages	84	7.017
Male	Health and beauty	88	6.899

Gender <chr>	Productline <chr>	counts <int>	Averagerating <dbl>
Male	Home and lifestyle	81	6.910
Male	Sports and travel	78	7.085





NOTE : A lot of indepth analysis related to exact time of purchases wrt customer type, product category, ratings by categories in each branch etc can be explored heavily further.