MATERIALS_DATA DATASET

Loading the dataset:

- The Materials data dataset is loaded into a data frame using R
- We use head function to retrieve the number of rows in the data frame

```
R 4.1.2 · ~/
> df_data<-read.csv("C:/users/Puddi/Desktop/Masters_UTA/MS_Sem2/CSE-5334/Assignment1/patel-1/patel/patel/marketing_data.csv")
>
> head(df_data, n = 5L)
  i..ID Year_Birth  Education Marital_Status      Income Kidhome Teenhome Dt_Customer Recency MntWines MntFruits MntMeatProducts
1  1826       1970 Graduation       Divorced $84,835.00        0        0     6/16/14       0      189       104             379
2     1       1961 Graduation         Single $57,091.00        0        0     6/15/14       0      464         5              64
3 10476       1958 Graduation        Married $67,267.00        0        1     5/13/14       0      134        11              59
4  1386       1967 Graduation       Together $32,474.00        1        1     5/11/14       0       10         0               1
5  5371       1989 Graduation         Single $21,474.00        1        0      4/8/14       0        6        16              24
  MntFishProducts MntSweetProducts MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases NumStorePurchases NumWebVisitsMonth
1             111              189          218                 1                4                   4                 6                 1
2               7                0           37                 1                7                   3                 7                 5
3              15                2           30                 1                3                   2                 5                 2
4               0                0            0                 1                1                   0                 2                 7
5              11                0           34                 2                3                   1                 2                 7
  AcceptedCmp3 AcceptedCmp4 AcceptedCmp5 AcceptedCmp1 AcceptedCmp2 Response Complain Country
1            0            0            0            0            0        1        0      SP
2            0            0            0            0            1        1        0      CA
3            0            0            0            0            0        0        0      US
4            0            0            0            0            0        0        0     AUS
5            1            0            0            0            0        1        0      SP
>
```

TASK 1 - Statistical Exploratory Data Analysis

Task1-a: Printing the details of the data frame

- We use data.frame to print all the details of the data frame in R

```
> print.data.frame(df_data)
   i..ID Year_Birth  Education Marital_Status      Income Kidhome Teenhome Dt_Customer Recency MntWines MntFruits MntMeatProducts
1   1826       1970 Graduation       Divorced $84,835.00        0        0     6/16/14       0      189       104             379
2      1       1961 Graduation         Single $57,091.00        0        0     6/15/14       0      464         5              64
3  10476       1958 Graduation        Married $67,267.00        0        1     5/13/14       0      134        11              59
4   1386       1967 Graduation       Together $32,474.00        1        1     5/11/14       0       10         0               1
5   5371       1989 Graduation         Single $21,474.00        1        0      4/8/14       0        6        16              24
6   7348       1958        PhD         Single $71,691.00        0        0     3/17/14       0      336       130             411
7   4073       1954   2n Cycle        Married $63,564.00        0        0     1/29/14       0      769        80             252
8   1991       1967 Graduation       Together $44,931.00        0        1     1/18/14       0       78         0              11
9   4047       1954        PhD        Married $65,324.00        0        1     1/11/14       0      384         0             102
10  9477       1954        PhD        Married $65,324.00        0        1     1/11/14       0      384         0             102
11  2079       1947   2n Cycle        Married $81,044.00        0        0    12/27/13       0      450        26             535
12  5642       1979     Master       Together $62,499.00        1        0    12/9/13        0      140         4              61
13 10530       1959        PhD          Widow $67,786.00        0        0    12/7/13        0      431        82             441
14  2964       1981 Graduation        Married $26,872.00        0        0    10/16/13       0        3        10               8
15 10311       1969 Graduation        Married  $4,428.00        0        1     10/5/13       0       16         4              12
16   837       1977 Graduation        Married $54,809.00        1        1     9/11/13       0       63         6              57
17 10521       1977 Graduation        Married $54,809.00        1        1     9/11/13       0       63         6              57
18 10175       1958        PhD       Divorced $32,173.00        0        1      8/1/13       0       18         0               2
19  1473       1960   2n Cycle         Single $47,823.00        0        1     7/23/13       0       53         1               5
20  2795       1958     Master         Single $30,523.00        2        1      7/1/13       0        5         0               3
21  2285       1954     Master       Together $36,634.00        0        1     5/28/13       0      213         9              76
22   115       1966     Master         Single $43,456.00        0        1     3/26/13       0      275        11              68
23 10470       1979     Master        Married $40,662.00        1        0     3/15/13       0       40         2              23
24  4065       1976        PhD        Married $49,544.00        1        0     2/12/13       0      308         0              73
25 10968       1969 Graduation         Single $57,731.00        0        1    11/23/12       0      266        21             300
26  5985       1965     Master         Single $33,168.00        0        1    10/13/12       0       80         1              37
27  5430       1956 Graduation       Together $54,450.00        1        1     9/14/12       0      454         0             171
28  8432       1956 Graduation       Together $54,450.00        1        1     9/14/12       0      454         0             171
29   453       1956        PhD          Widow $35,340.00        1        1     6/29/14       1       27         0              12
30  9687       1975 Graduation         Single $73,170.00        0        0     5/31/14       1      184       174             256
31  8890       1971        PhD       Divorced $65,808.00        1        1     5/30/14       1      155         7              80
32  9264       1986 Graduation        Married $79,529.00        0        0     4/27/14       1      423        42             706
33  5824       1972        PhD       Together $34,578.00        2        1     4/11/14       1        7         0               1
34  5794       1974        PhD        Married $46,374.00        0        1     3/17/14       1      408         0              21
```

	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth
1	111	189	218	1	4	4	6	1
2	7	0	37	1	7	3	7	5
3	15	2	30	1	3	2	5	2
4	0	0	0	1	1	0	2	7
5	11	0	34	2	3	1	2	7
6	240	32	43	1	4	7	5	2
7	15	34	65	1	10	10	7	6
8	0	0	7	1	2	1	3	5
9	21	32	5	3	6	2	9	4
10	21	32	5	3	6	2	9	4
11	73	98	26	1	5	6	10	1
12	0	13	4	2	3	1	6	4
13	80	20	102	1	3	6	6	1
14	3	16	32	1	1	1	2	6
15	2	4	321	0	25	0	0	1
16	13	13	22	4	2	1	5	4
17	13	13	22	4	2	1	5	4
18	0	0	2	1	1	0	3	4
19	2	1	10	2	2	0	3	8
20	0	0	5	1	1	0	2	7
21	4	3	30	3	5	2	5	7
22	25	7	7	3	5	1	8	5
23	0	4	23	2	2	1	3	4
24	0	0	23	2	5	1	8	7
25	65	8	44	4	8	8	6	6
26	0	1	3	3	2	1	4	7
27	8	19	32	12	9	2	8	8
28	8	19	32	12	9	2	8	8
29	0	1	5	2	2	0	3	5
30	50	30	32	1	5	4	6	2
31	13	7	10	3	5	1	5	6
32	73	197	197	1	4	8	9	2
33	0	0	0	1	1	0	2	6
34	0	0	17	3	7	1	7	8
35	0	14	7	1	2	0	3	7

	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Response	Complain	Country
1	0	0	0	0	0	1	0	SP
2	0	0	0	0	1	1	0	CA
3	0	0	0	0	0	0	0	US
4	0	0	0	0	0	0	0	AUS
5	1	0	0	0	0	1	0	SP
6	0	0	0	0	0	1	0	SP
7	1	0	0	0	0	1	0	GER
8	0	0	0	0	0	0	0	SP
9	0	0	0	0	0	0	0	US
10	0	0	0	0	0	0	0	IND
11	0	0	0	0	0	0	0	US
12	0	0	0	0	0	0	0	SP
13	0	0	0	0	0	1	0	IND
14	0	0	0	0	0	0	0	CA
15	0	0	0	0	0	0	0	SP
16	0	0	0	0	0	0	0	SP
17	0	0	0	0	0	1	0	SP
18	0	0	0	0	0	0	0	SP
19	0	0	0	0	0	0	0	CA
20	0	0	0	0	0	0	0	CA
21	0	0	0	0	0	0	0	SA
22	0	0	0	0	0	0	0	IND
23	0	0	0	0	0	0	0	GER
24	0	0	0	0	0	0	0	SP
25	0	0	0	0	0	0	0	IND
26	0	0	0	0	0	0	0	SP
27	0	0	0	0	0	0	0	SP
28	0	0	0	0	0	0	0	SP
29	0	0	0	0	0	0	0	SP
30	0	0	0	0	0	0	0	CA
31	0	0	0	0	0	0	0	SP
32	0	0	0	0	0	0	0	CA
33	0	0	0	0	0	0	0	AUS
34	0	1	0	1	0	1	0	IND
35	0	0	0	0	0	0	0	SP

[reached 'max' / getOption("max.print") -- omitted 2205 rows]

Task 1-b: Finding the number of rows and columns in dataset

- To find the length of the rows and columns we use nrow and ncol in R

```
> nrow(df_data)
[1] 2240
> ncol(df_data)
[1] 28
>
```

Task 1-c: descriptive detail of a 'Year_Birth' and 'MntMeatProducts' column in dataset

- We have summary function in R which will provide all the descriptive details of the particular column in a data frame

```
> summary(df_data$Year_Birth)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1893    1959    1970    1969    1977    1996
> summary(df_data$MntMeatProducts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    16.0    67.0   166.9   232.0  1725.0
>
```

Task 1-d: Finding all the unique values for a column year_Birth and its respective length.

- We have 'unique' function in R which will provide all the unique values
- To calculate the respective length, we use length function

```
> (unique(d))
 [1] 1970 1961 1958 1967 1989 1954 1947 1979 1959 1981 1969 1977 1960 1966 1976 1965 1956 1975 1971 1986 1972 1974 1990 1987 1984 1968 1955
[28] 1983 1973 1978 1952 1962 1964 1982 1963 1957 1980 1945 1949 1948 1953 1946 1985 1992 1944 1951 1988 1950 1994 1993 1991 1893 1996 1995
[55] 1899 1943 1941 1940 1900
> length(unique(d))
[1] 59
>
```

Task 2-a: Data whose income is more than 100K

- We have used subset function to find the IDs whose income is more than 100000 lakh and used nrow for the count of the rows

```
> salary_slab<-subset(df_data1, df_data1$Income > 100000)
> (salary_slab)
     ï..ID Year_Birth  Education Marital_Status Income Kidhome Teenhome Dt_Customer Recency MntWines MntFruits MntMeatProducts
143  10089       1974 Graduation       Divorced 102692       0        0     4/5/13       5      168       148             444
211   4619       1945        PhD         Single 113734       0        0     5/28/14      9        6         2               3
326   4931       1977 Graduation       Together 157146       0        0     4/29/13     13        1         0            1725
498   1501       1982        PhD        Married 160803       0        0      8/4/12     21       55        16            1622
528   9432       1977 Graduation       Together 666666       1        0      6/2/13     23        9        14              18
732   1503       1976        PhD       Together 162397       1        1      6/3/13     31       85         1              16
833   4611       1970 Graduation       Together 105471       0        0     1/21/13     36     1009       181             104
854   5336       1971     Master       Together 157733       1        0      6/4/13     37       39         1               9
1245  2798       1977        PhD       Together 102160       0        0     11/2/12     54      763        29             138
1565  7215       1983 Graduation         Single 101970       0        0     3/12/13     69      722        27             102
1827  5555       1975 Graduation       Divorced 153924       0        0      2/7/14     81        1         1               1
1926 11181       1949        PhD        Married 156924       0        0     8/29/13     85        2         1               2
2205  8475       1973        PhD        Married 157243       0        1      3/1/14     98       20         2            1582
     MntFishProducts MntSweetProducts MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases NumStorePurchases NumWebVisitsMonth
143               32              172         148                 1               6                   9                13                 2
211                1              262           3                 0              27                   0                 0                 1
326                2                1           1                 0               0                  28                 0                 1
498               17                3           4                15               0                  28                 1                 0
528                8                1          12                 4               3                   1                 3                 6
732                2                1           2                 0               0                   0                 1                 1
833              202               21         207                 0               9                   8                13                 3
854                2                0           8                 0               1                   0                 1                 1
1245              76              176          58                 0               7                   9                10                 4
1565              44               72         168                 0               6                   8                13                 2
1827               1                1           1                 0               0                   0                 0                 0
1926               1                1           1                 0               0                   0                 0                 0
2205               1                2           1                15               0                  22                 0                 0
     AcceptedCmp3 AcceptedCmp4 AcceptedCmp5 AcceptedCmp1 AcceptedCmp2 Response Complain Country
143             0            1            1            1            1        1        0      SA
211             0            0            0            0            0        0        0      SP
326             0            0            0            0            0        0        0      SA
498             0            0            0            0            0        0        0      US
528             0            0            0            0            0        0        0      SA
732             0            0            0            0            0        0        0      SP
833             0            0            1            1            0        1        0      SP
854             0            0            0            0            0        0        0      SP
1245            0            1            1            1            0        1        0      SA
1565            0            1            1            1            0        1        0      CA
1827            0            0            0            0            0        0        0      SP
1926            0            0            0            0            0        0        0      CA
2205            0            0            0            0            0        0        0     IND
> nrow(salary_slab)
[1] 13
>
```

Task 2-b: Number of customers born between 1990 and 2000

- We have created a new subset from the existing data frame because we preprocess the data and use it for this new dataframe
- Now we use nrow to get the count of the number of customers born between year 1990 and 2000

```
> data_interest <- subset(df_data1, df_data1$Year_Birth >=1960 & df_data1$Year_Birth <=1970)
> nrow(data_interest)
[1] 583
>
```

Task 2C: Top 10 IDs with the highest Income

- To find the top 10 IDs with highest income we can sort the data frame with the highest income to the descending order
- This will retrive the top 10 values of the data frame
- We have used Head function to retrive the top 10 IDs with highest income

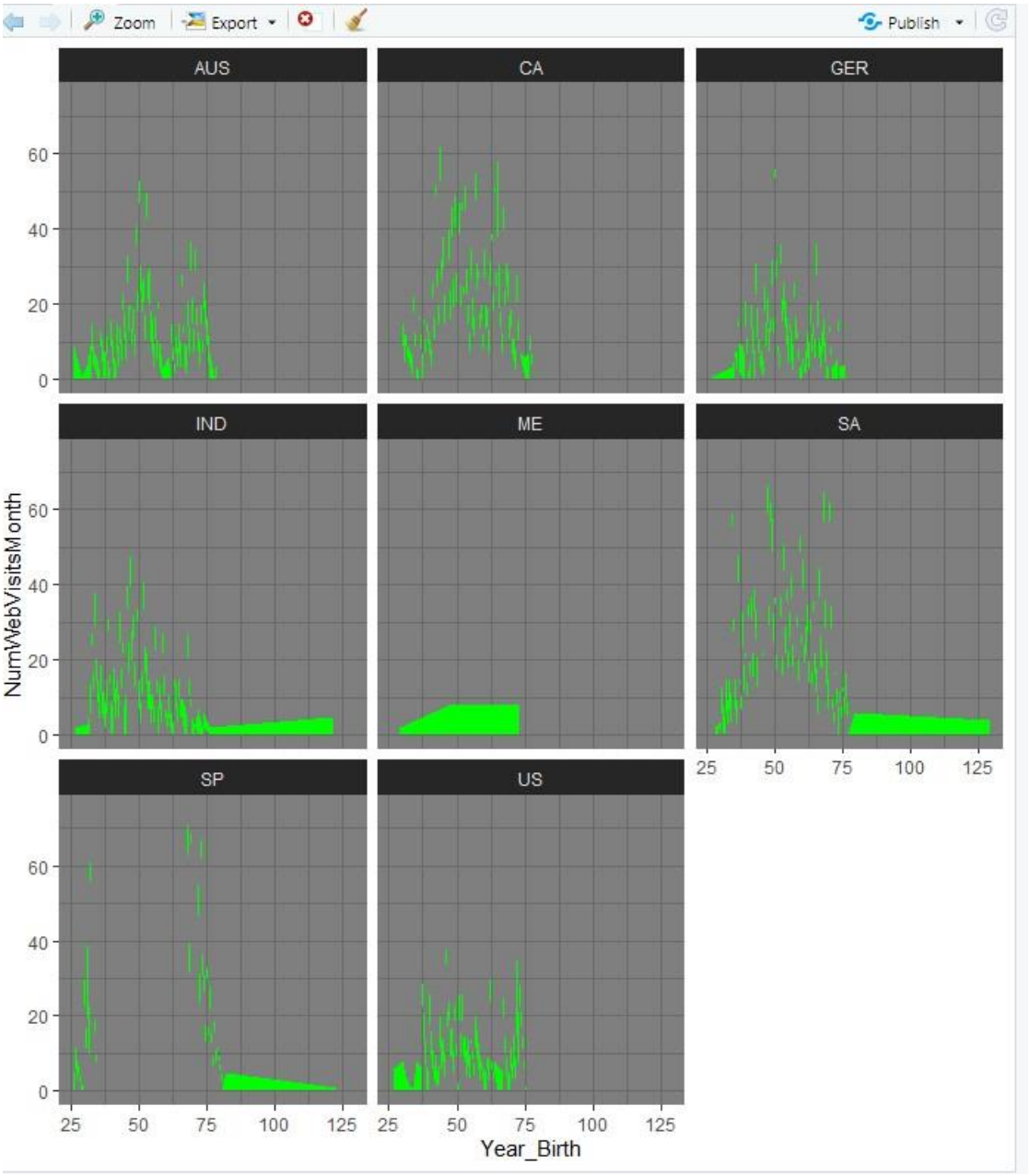
```
> task2c<-df_data1[order(df_data1$Income, decreasing = TRUE), ]
> head(select(task2c,1,5), n = 10L)
     i..ID Income
528   9432 666666
732   1503 162397
498   1501 160803
854   5336 157733
2205  8475 157243
326   4931 157146
1926 11181 156924
1827  5555 153924
211   4619 113734
833   4611 105471
```

##TASK 3: VISUALIZATION

Task 3-a: Plotting the comparison of number of web visits with Year of Birth

- ggplot is a package in R which helps to plot the data in the data frame through the axes
- We use geom_area function to plot the comparision of number of web visits with Year of Birth with colour ‘fill’
- facet_wrap() as a ribbon of plots that arranges panels into rows and columns and chooses a layout that best fits the number of panels.
- We use this function for each country in the data frame and compare each country

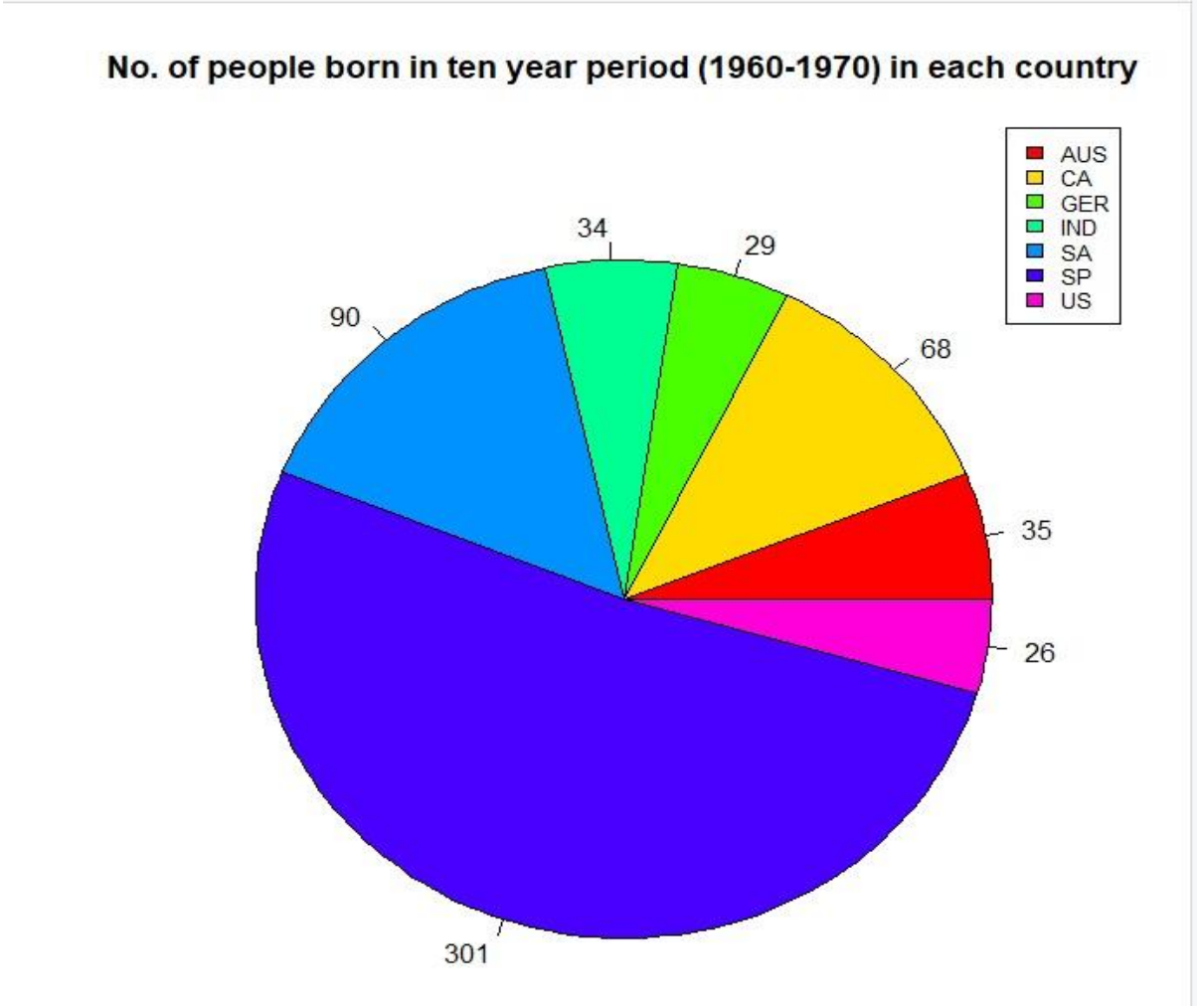
```
> task3a<-ggplot(df_data2, aes(x = Year_Birth, y = NumWebVisitsMonth)) +
+ geom_area(fill='#00FF00', alpha=2) +
+ facet_wrap(~ Country)
> task3a+ylim(0,75)+ theme_dark()
>
```



Task 3-b: Pie chart that shows the number of people born between 1960 and 1970 in each country.

- Ggplot is used to draw a pie chart for number of people born between 1960 and 1970
- We use the column ‘year_Birth’ to count the values using the summarise function
- The title is changed using the main() function
- We use geom_area function to plot the comparision of number of births with Year of Birth in each country
- We use ggplot to create a pie chart with size using labels and the total we counted earlier using the summarise function

```
> task3a<-ggplot(df_data2, aes(x = Year_Birth, y = NumWebVisitsMonth)) +
+ geom_area(fill='#00FF00', alpha=2) +
+ facet_wrap(~ Country)
> task3a+ylim(0,75)+ theme_dark()
> data_interest1 <- subset(df_data1, df_data1$Year_Birth >=1960 & df_data1$Year_Birth <=1970)
> gfh<-group_by(data_interest ,Country)%>%summarise(Total=n())
> ##head(gfh, n=15L)
>
> pie(gfh$Total , labels = paste0(gfh$Total),
+     main = "No. of people born in ten year period (1960-1970) in each country", col = rainbow(length(gfh$Total)))
> legend("topright", c(gfh$Country), cex = 0.8,
+        fill = rainbow(length(gfh$Total)))
>
```

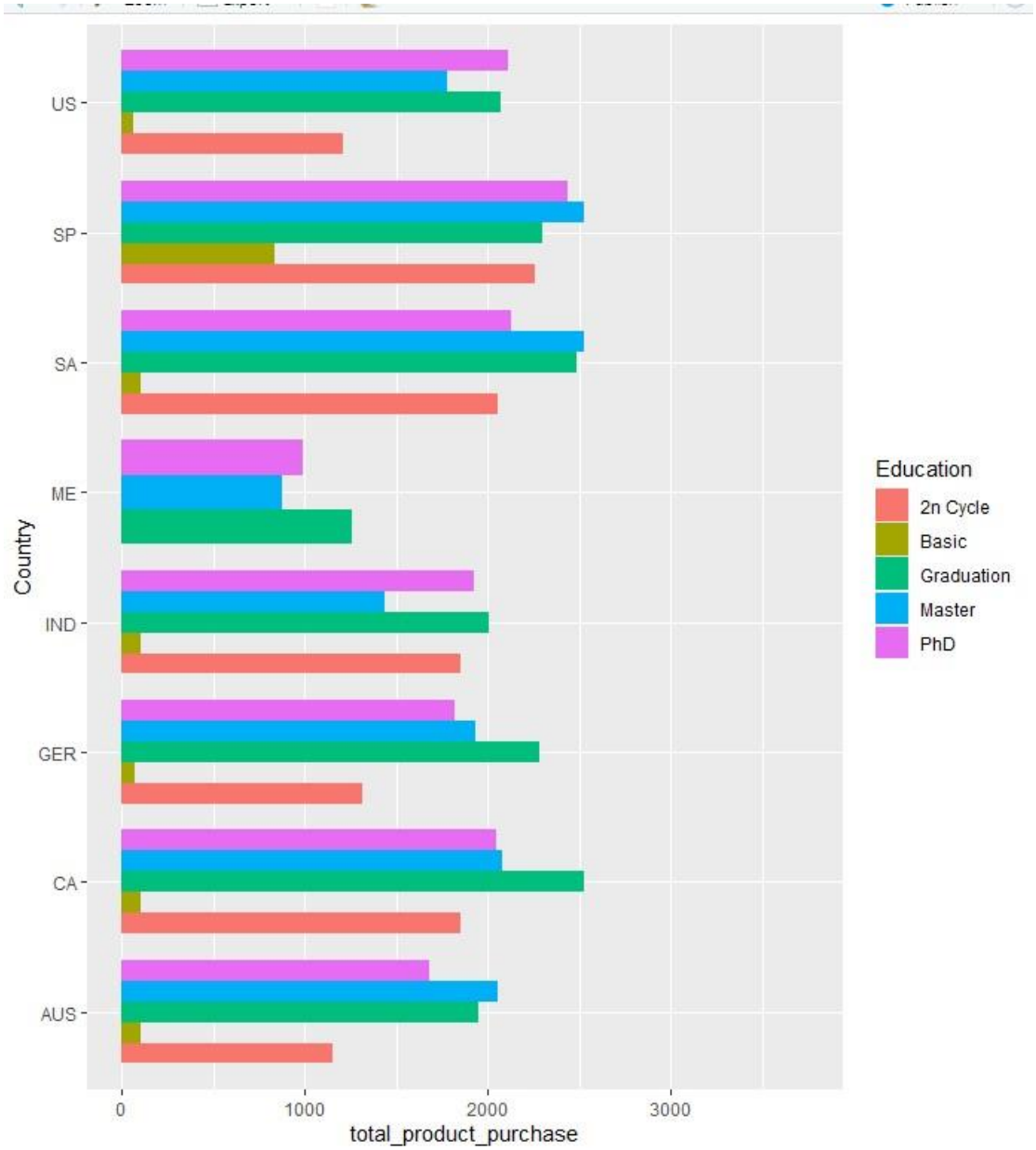


Task 4: Pattern Visualization

- To find an interesting pattern we have used columns ‘’Country” and “education” to find the number of products purchased with their education level for each country
- We calculated the total using summarise function and grouped with the dataframe country and education
- Group_by is used to group the data with country and education
- ggplot is used to map the data and geom_bar is used to position the data in the graph
- From the graph we could see the number of products purchased for each country based on their education

Code:

```
> asd1<-group_by(df_data2,Country,Education)%>%summarise(total_product_purchase)
`summarise()` has grouped output by 'Country', 'Education'. You can override using the `.groups` argument.
> task4<-asd1%>%ggplot(mapping=aes(y=Country,x=total_product_purchase,fill=Education))+
+   geom_bar(stat='identity',position = 'dodge', width=.8)
>
> task4+theme_gray()+xlim(0,3750)
>
```



References:

Most of the basic functions has been retrieved using R Studio Console http://cran.us.r-project.org

https://rstudio-pubs-static.s3.amazonaws.com/154893_534f69f0c78c46ea80aca8282fbb38e9.html