Nandini Mehta

3CS2

Roll No-102067009

Data Science Project- StartUp Expansion visualisation using tableau and R

Dataset used-

https://drive.google.com/drive/folders/1_s1GTLlbekjkPXxCEAmrPKJrti_3FCLd?usp=sharing

```
> #Loading the dataset in R
> getwd()
[1] "C:/Users/Nandini/OneDrive/Documents"
> df = read.csv("StartupExpansion.csv")
> df2 = read.csv("Additional.csv")
> summary(df)
    Store.ID          City              State           Sales_Region
 Min.   :  1.00   Length:166         Length:166         Length:166
 1st Qu.: 39.25   Class :character   Class :character   Class :character
 Median : 74.50   Mode  :character   Mode  :character   Mode  :character
 Mean   : 75.69
 3rd Qu.:112.75
 Max.   :150.00
 New.Expansion      Marketing_Spend      Revenue              Age
 Length:166         Length:166         Length:166         Length:166
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
> colnames(df)
[1] "Store.ID"        "City"            "State"           "Sales_Region"
[5] "New.Expansion"   "Marketing_Spend" "Revenue"         "Age"
```

```
> colnames(df)
[1] "Store.ID"        "City"            "State"           "Sales_Region"
[5] "New.Expansion"   "Marketing_Spend" "Revenue"         "Age"
> head(df)
  Store.ID          City      State Sales_Region New.Expansion Marketing_Spend
1        1        Peoria    Arizona     Region 2           Old          $2,601
2        2       Midland      Texas     Region 2           Old          $2,727
3        3       Spokane Washington     Region 2           Old          $2,768
4        4        Denton      Texas     Region 2           Old          $2,759
5        5 Overland Park     Kansas     Region 2           Old          $2,869
6        6       Yonkers   New York     Region 1           Old          $3,080
  Revenue            Age
1 $48,610 16 to 19 years
2 $45,689 20 to 24 years
3 $49,554 25 to 34 years
4 $38,284 35 to 44 years
5 $59,887 45 to 54 years
6 $53,827 55 to 64 years
>
```

#Query 1

List columns with Na values and remove them

```
> df$Marketing_Spend = trimws(gsub("\\$", "", df$Marketing_Spend))
> df$Revenue = gsub("\\$", "", df$Revenue)
> #Removing NA values
> list_na <- colnames(df)[ apply(df, 2, anyNA) ]
> list_na
[1] "City"          "State"         "Sales_Region"  "New.Expansion"
[5] "Age"
> library(dplyr)


Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> df_drop <-df %>%
+   na.omit()
>   dim(df_drop)
[1] 158   8
```

#Query 2

Replace values other than "Region1, Region2" in the "Sales Region" column

```
> #Replace values other than "Region1, Region2" in the "Sales Region" column
> v <- c("Region 1","Region 2")
> getmode <- function(v) {
+   uniqv <- unique(v)
+   uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
> df_drop$Sales_Region[!df_drop$Sales_Region %in% v] <- getmode(df_drop$Sales_Region)
> df_drop$Sales_Region
  [1] "Region 2" "Region 2" "Region 2" "Region 2" "Region 2" "Region 1"
  [7] "Region 1" "Region 2" "Region 2" "Region 1" "Region 1" "Region 2"
 [13] "Region 2" "Region 2" "Region 2" "Region 2" "Region 2" "Region 1"
 [19] "Region 1" "Region 2" "Region 2" "Region 1" "Region 2" "Region 2"
 [25] "Region 2" "Region 2" "Region 2" "Region 1" "Region 2" "Region 2"
 [31] "Region 1" "Region 1" "Region 2" "Region 2" "Region 1" "Region 2"
 [37] "Region 1" "Region 2" "Region 1" "Region 1" "Region 2" "Region 2"
 [43] "Region 2" "Region 1" "Region 1" "Region 1" "Region 2" "Region 1"
 [49] "Region 2" "Region 2" "Region 2" "Region 2" "Region 2" "Region 2"
 [55] "Region 2" "Region 2" "Region 2" "Region 1" "Region 2" "Region 1"
 [61] "Region 2" "Region 2" "Region 2" "Region 2" "Region 1" "Region 1"
 [67] "Region 2" "Region 1" "Region 1" "Region 2" "Region 2" "Region 2"
```

#Query 3

Replace rows with negative values in the "Marketing Spend" column

```
> #Replace rows with negative values in the "Marketing Spend" column
> df_drop$Marketing_Spend=as.numeric(df_drop$Marketing_Spend)
```

```
> df_drop$Marketing_Spend
  [1] "2,601"  "2,727"  "2,768"  "2,759"  "2,869"  "3,080"  "3,110"  "2,593"
  [9] "-2,593" "2,675"  "2,984"  "-2,768" "2,759"  "2,541"  "2,651"  "2,895"
 [17] "-3,675" "-2,648" "3,466"  "2,686"  "2,795"  "2,737"  "3,085"  "2,894"
 [25] "2,765"  "2,521"  "3,077"  "3,287"  "3,679"  "2,918"  "2,420"  "2,557"
 [33] "3,620"  "2,483"  "3,131"  "3,083"  "2,808"  "2,984"  "2,484"  "3,335"
 [41] "3,127"  "2,904"  "2,318"  "3,488"  "2,886"  "2,373"  "3,262"  "2,758"
 [49] "2,607"  "3,146"  "3,399"  "2,790"  "3,220"  "2,344"  "2,939"  "2,648"
 [57] "3,082"  "2,338"  "3,246"  "2,374"  "3,129"  "2,939"  "2,642"  "3,082"
 [65] "2,871"  "3,392"  "3,228"  "3,175"  "2,874"  "2,792"  "3,924"  "2,857"
 [73] "-2,857" "2,440"  "2,950"  "2,462"  "3,084"  "3,177"  "3,003"  "3,697"
 [81] "2,857"  "3,168"  "2,943"  "2,777"  "3,329"  "2,854"  "2,555"  "3,162"
 [89] "2,928"  "2,275"  "3,285"  "3,217"  "3,115"  "2,779"  "3,277"  "2,745"
 [97] "3,984"  "3,262"  "2,718"  "2,552"  "3,072"  "2,676"  "2,553"  "2,685"
[105] "2,990"  "2,805"  "2,846"  "2,699"  "2,901"  "3,015"  "2,160"  "3,079"
[113] "2,104"  "2,962"  "2,830"  "3,228"  "3,005"  "3,020"  "2,947"  "3,131"
[121] "2,901"  "2,926"  "3,086"  "2,754"  "2,282"  "3,319"  "3,081"  "2,542"
[129] "2,527"  "2,967"  "2,357"  "2,838"  "2,914"  "2,854"  "2,759"  "2,553"
[137] "2,877"  "3,507"  "2,608"  "2,028"  "2,995"  "3,586"  "3,067"  "1,811"
[145] "2,736"  "3,112"  "2,603"  "3,191"  "3,587"  "2,911"  "3,279"  "2,945"
```

#Query 4

Replace rows with negative values in the "Revenue" column

```
> #Replace rows with negative values in the "Revenue" column
> df_drop$Revenue=as.numeric(df_drop$Revenue)
```

```
> df_drop$Revenue
  [1] "48,610"  "45,689"  "49,554"  "38,284"  "59,887"  "53,827"  "60,338"
  [8] "19,569"  "19,569"  "59,840"  "64,906"  "49,554"  "-38,284" "16,860"
 [15] "21,988"  "19,888"  "63,148"  "43,377"  "54,701"  "18,471"  "16,690"
 [22] "47,729"  "63,027"  "43,183"  "19,120"  "38,178"  "56,836"  "52,114"
 [29] "20,123"  "49,856"  "55,790"  "45,017"  "56,921"  "39,744"  "22,972"
 [36] "22,680"  "65,475"  "21,718"  "34,829"  "59,283"  "20,057"  "20,455"
 [43] "64,302"  "49,506"  "52,250"  "41,313"  "-1245"   "57,625"  "16,029"
 [50] "49,191"  "59,870"  "48,254"  "43,397"  "16,372"  "50,233"  "43,377"
 [57] "41,460"  "50,364"  "44,223"  "17,006"  "15,562"  "21,824"  "46,490"
 [64] "41,460"  "38,782"  "19,350"  "16,652"  "23,764"  "42,803"  "41,110"
 [71] "19,448"  "18,838"  "52,078"  "68,828"  "19,529"  "29,008"  "55,684"
 [78] "45,418"  "18,297"  "18,966"  "18,838"  "52,078"  "25,321"  "40,545"
 [85] "58,951"  "40,921"  "49,609"  "45,550"  "19,563"  "41,361"  "57,530"
 [92] "54,768"  "18,754"  "21,286"  "48,796"  "19,708"  "56,089"  "19,752"
 [99] "20,949"  "45,666"  "40,779"  "47,482"  "18,215"  "50,650"  "48,933"
[106] "42,026"  "15,735"  "39,856"  "20,669"  "20,541"  "33,647"  "41,319"
[113] "45,632"  "43,773"  "36,821"  "22,824"  "21,953"  "51,229"  "50,583"
[120] "50,895"  "46,503"  "18,089"  "56,504"  "44,635"  "43,924"  "19,811"
[127] "56,140"  "40,462"  "54,145"  "51,882"  "35,022"  "49,290"  "47,108"
```

#Query 5

To remove outliers in the "age" column

```
> #To remove outliers in the "age" column
> x <- round(mean(df2$Employed))
> df2$Employed[df2$Employed > 200000] <- x
> df2$Employed
  [1]  91000 175000 194000 112347 112347 101000  33000  38000  90000 142000
 [11] 180000 157000  71000  82000  90000 154000 176000 184000 186000  98000
 [21]  38000  54000 114000 119000 170000 173000  94000  85000 114000 137000
 [31] 195000 112347 112347 115000  32000  53000 113000 102000 141000 152000
 [41] 101000  72000  86000 125000 191000 112347 181000  93000  31000  53000
 [51] 114000 151000 172000 160000  90000  80000  87000 106000 183000 173000
 [61] 159000  87000  22000  44000  99000 149000 144000 178000  74000  71000
 [71]  74000 109000 162000 146000 171000  64000  37000  47000  73000 114000
 [81] 114000 150000  58000  89000  79000 110000 139000 164000 181000  56000
 [91]  36000  39000  72000 156000 115000 123000  81000  86000  69000 142000
[101] 141000 157000 163000  82000  25000  47000  57000 151000 137000 130000
[111]  85000 101000  64000 127000 161000 152000 146000  83000  36000  41000
[121]  55000 163000 149000 152000  80000  96000  51000 102000 152000 137000
[131] 195000  90000  41000  48000  56000 150000 156000 120000  69000  94000
[141]  53000  97000 137000 151000 197000  82000  37000  28000  65000 140000
>
```

#Query 6

Randomly splitting the dataset

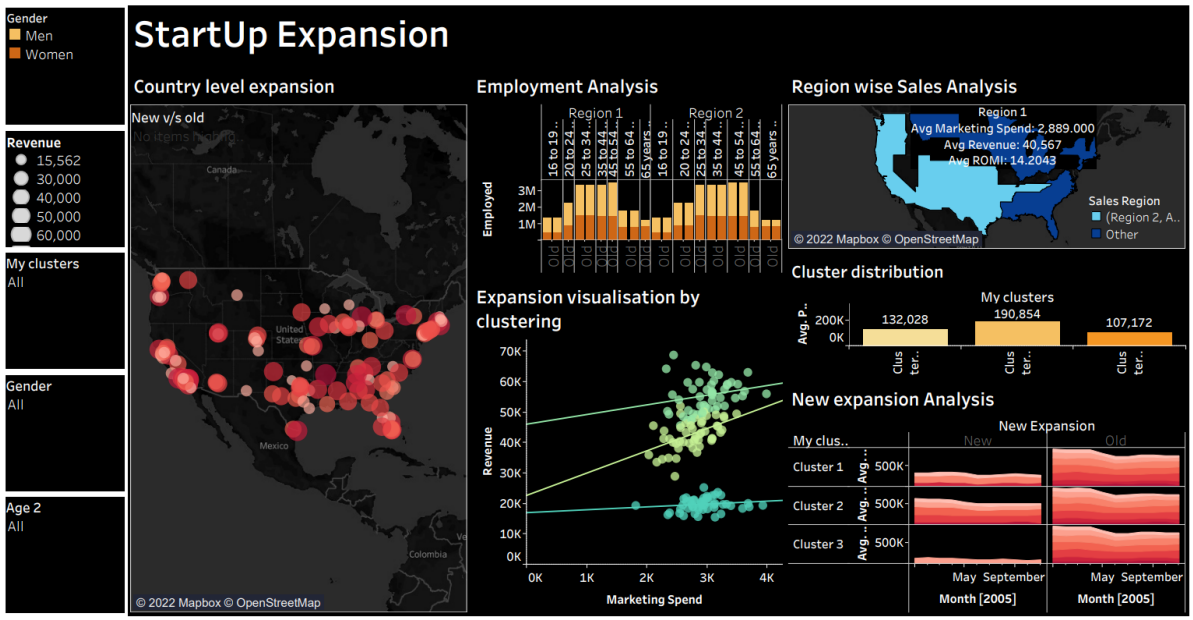
```
> #Randomly splitting the dataset
> id = sample(2, nrow(df_drop), replace = TRUE, prob = c(0.6, 0.4))
> grp1 = df_drop[id==1,]
> grp2 = df_drop[id==2,]
> dim(df_drop)
[1] 158   8
> dim(grp1)
[1] 94  8
> dim(grp2)
[1] 64  8
```

#Query 7

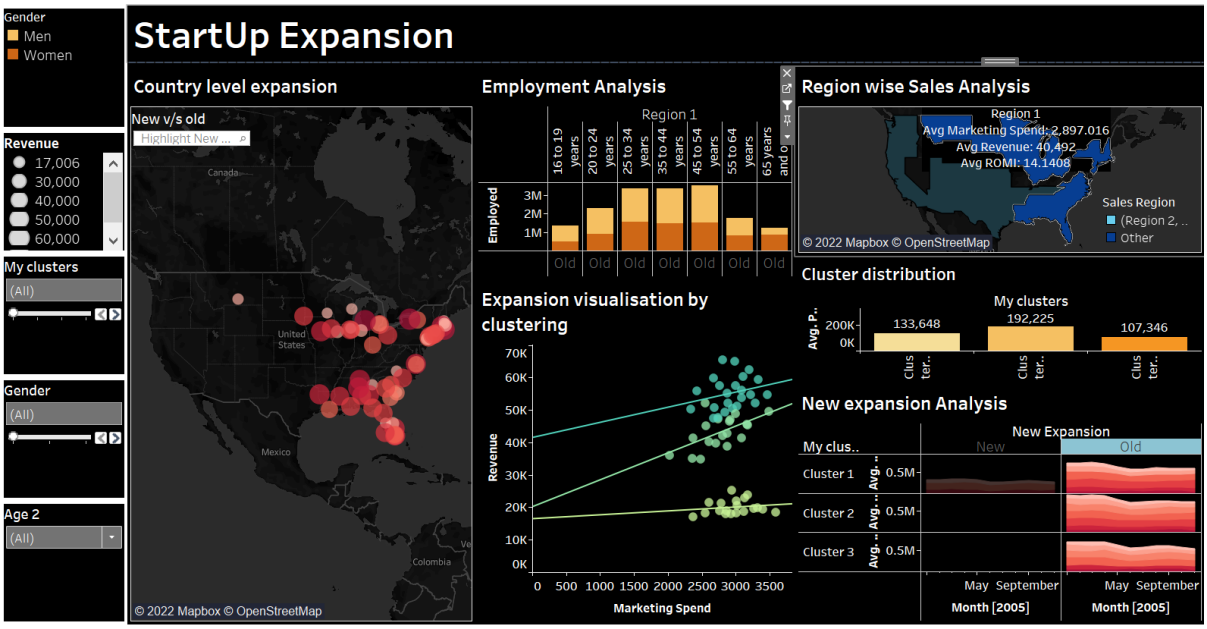
Merging groups and making a new file

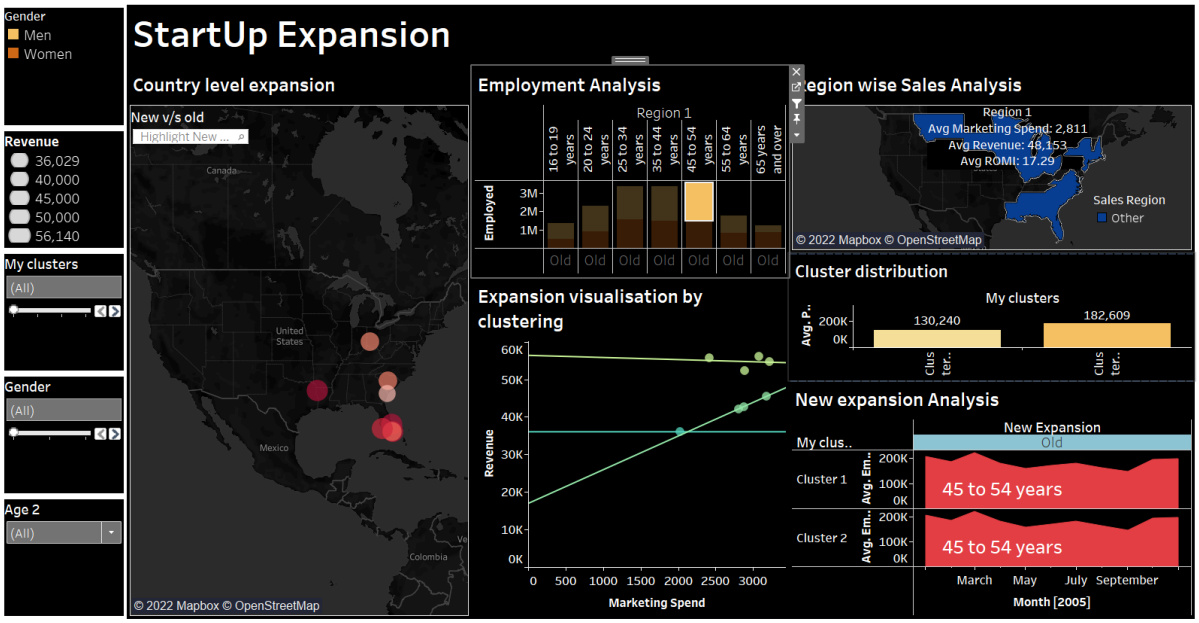
```
> #Merging grp1 and grp2 into one csv file
> dff <- rbind(grp1,grp2)
> write.csv(dff,file="Startup Expansion 2.csv")
```

Dashboard In Tableau

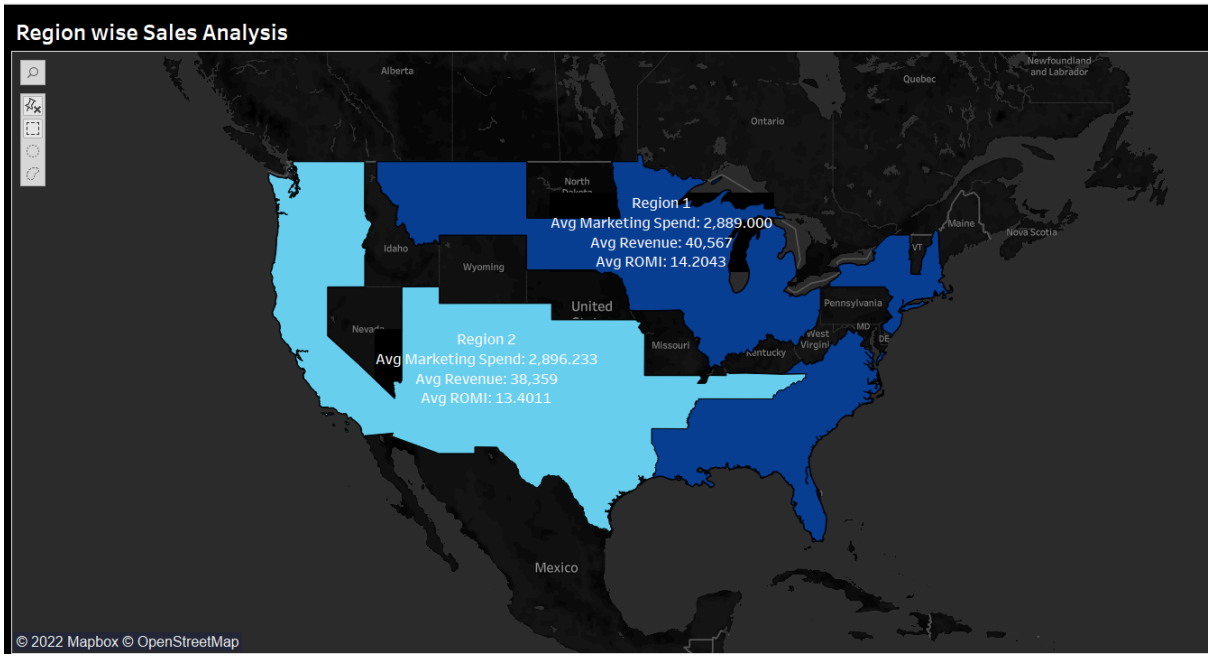


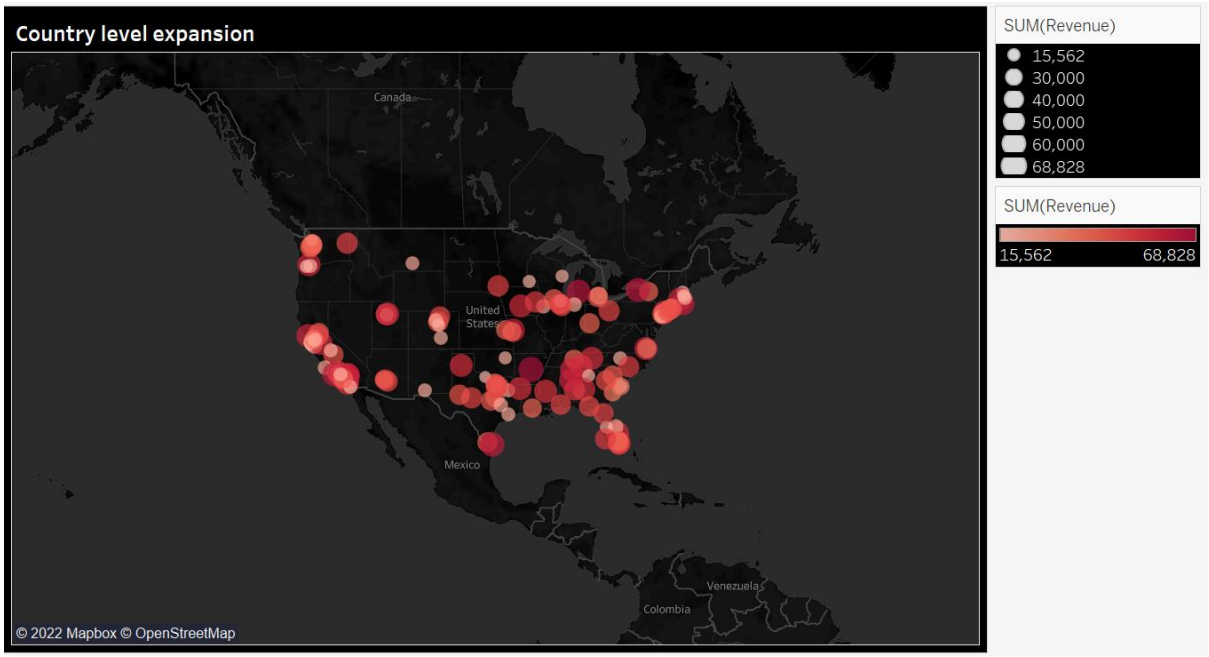
Samples-

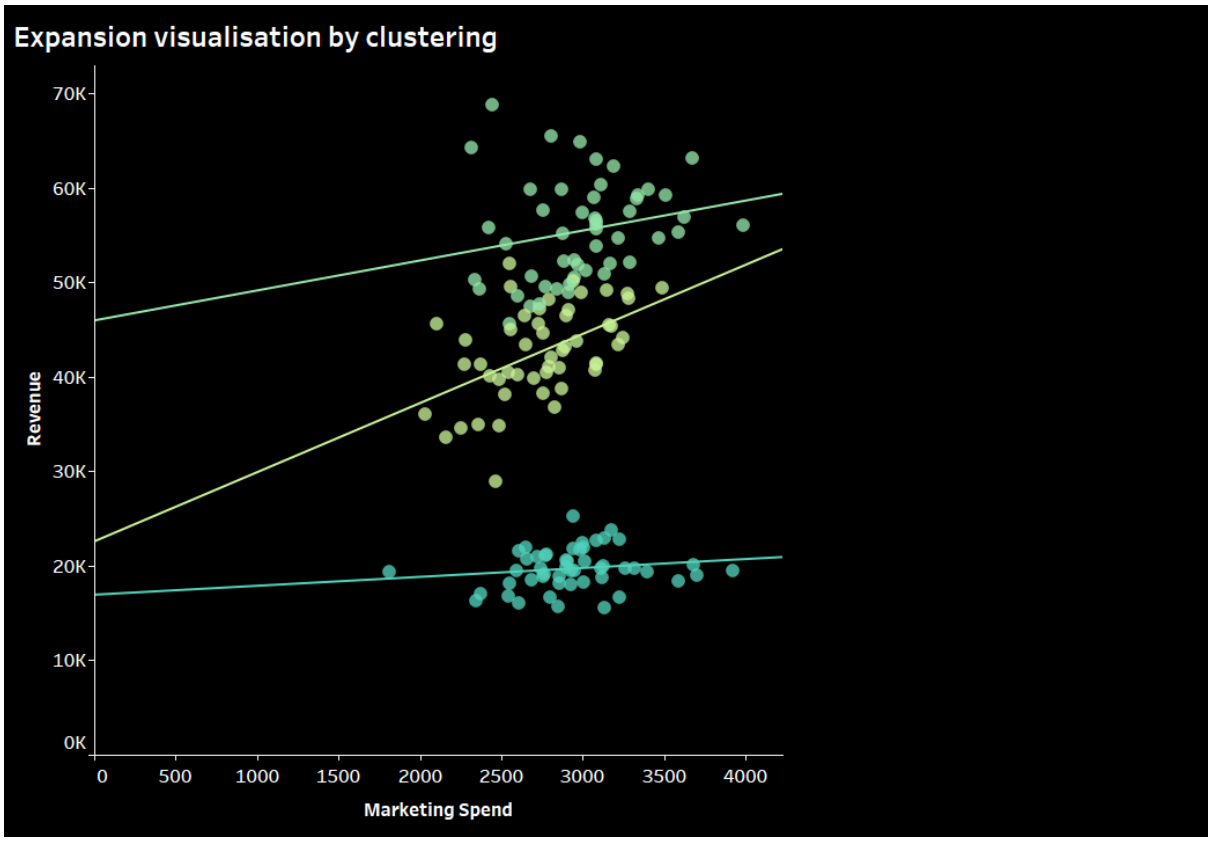



Gender
Men
Women
Revenue
36,029
40,000
45,000
50,000
56,140
My clusters
(All)
Gender
(All)
Age 2
(All)
StartUp Expansion
Country level expansion
New v/s old
Highlight New ...
Canada
United States
Mexico
Colombia
© 2022 Mapbox © OpenStreetMap
Employment Analysis
Region 1
16 to 19 years
20 to 24 years
25 to 34 years
35 to 44 years
45 to 54 years
55 to 64 years
65 years and over
Employed
3M
2M
1M
Old
Expansion visualisation by clustering
Revenue
60K
50K
40K
30K
20K
10K
0K
0
500
1000
1500
2000
2500
3000
Marketing Spend
egion wise Sales Analysis
Region 1
Avg Marketing Spend: 2,811
Avg Revenue: 48,153
Avg ROMI: 17.29
Sales Region
Other
© 2022 Mapbox © OpenStreetMap
Cluster distribution
My clusters
Avg. P..
200K
0K
130,240
182,609
Clus ter..
New expansion Analysis
My clus..
New Expansion
Old
Cluster 1
Cluster 2
Avg. Em..
200K
100K
0K
45 to 54 years
March
May
July
September
Month [2005]

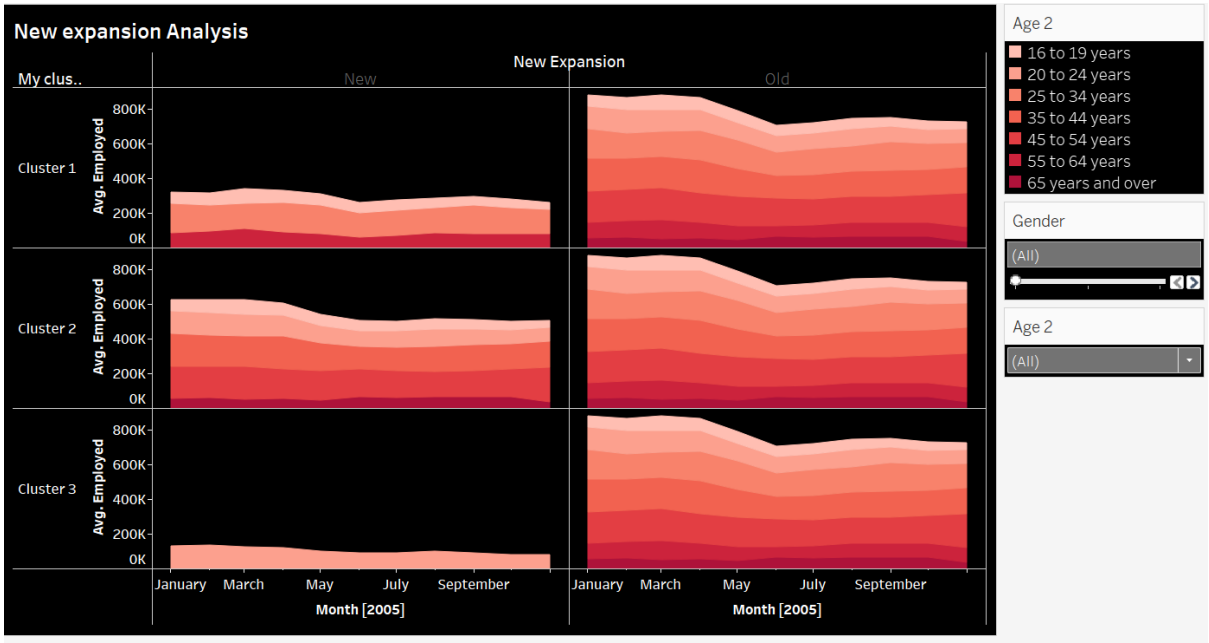
Worksheets-

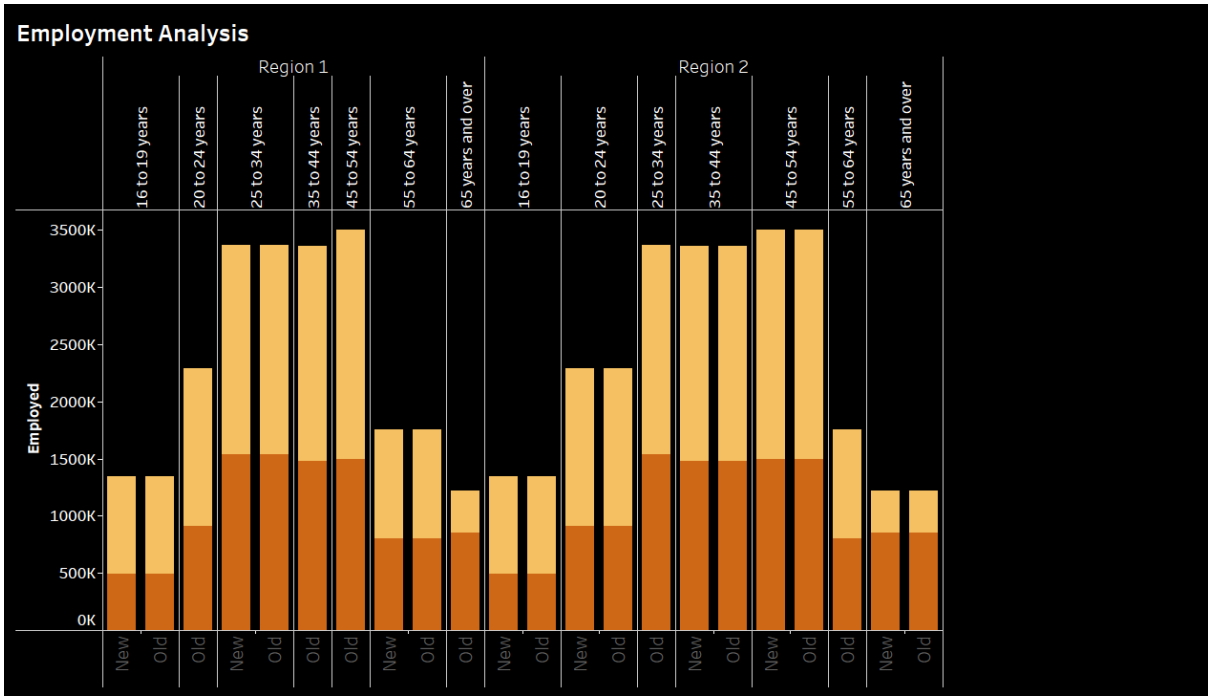

Region wise Sales Analysis
Alberta
Ontario
Quebec
Newfoundland and Labrador
Region 1
Avg Marketing Spend: 2,889.000
Avg Revenue: 40,567
Avg ROMI: 14.2043
Region 2
Avg Marketing Spend: 2,896.233
Avg Revenue: 38,359
Avg ROMI: 13.4011
Idaho
Wyoming
Nevada
Missouri
Pennsylvania
Maine
Nova Scotia
Mexico
© 2022 Mapbox © OpenStreetMap


Country level expansion
Canada
United States
Mexico
Venezuela
Colombia
© 2022 Mapbox © OpenStreetMap
SUM(Revenue)
15,562
30,000
40,000
50,000
60,000
68,828
SUM(Revenue)
15,562
68,828


Expansion visualisation by clustering
70K
60K
50K
40K
30K
20K
10K
0K
Revenue
0
500
1000
1500
2000
2500
3000
3500
4000
Marketing Spend


Cluster distribution
My clusters
Avg. Population
200K
180K
160K
140K
120K
100K
80K
60K
40K
20K
0K
132,028
190,854
107,172
Cluster 1
Cluster 2
Cluster 3
My clusters
Cluster 1
Cluster 2
Cluster 3
My clusters
(All)


New expansion Analysis
My clus..
New Expansion
New
Old
Cluster 1
Cluster 2
Cluster 3
Avg. Employed
800K
600K
400K
200K
0K
January
March
May
July
September
Month [2005]
Age 2
16 to 19 years
20 to 24 years
25 to 34 years
35 to 44 years
45 to 54 years
55 to 64 years
65 years and over
Gender
(All)
Age 2
(All)


Employment Analysis
Region 1
Region 2
16 to 19 years
20 to 24 years
25 to 34 years
35 to 44 years
45 to 54 years
55 to 64 years
65 years and over
Employed
3500K
3000K
2500K
2000K
1500K
1000K
500K
0K
New
Old