Segmenting the Smartwatch Market

Overview

SmartWatch performed cluster analysis to perform an industry segmentation in order to identify various subsamples of dentists who are homogeneous in their behaviors and preferences and markedly different from other subsamples. Use the dataset 'SmartWatch Case Data file.xlsm' to perform cluster analysis.

Solutions

1. **Understanding the data: Before embarking on the journey of modeling, it is important to cultivate a priliminary exploration of the data at hand; it lays the foundation for informed decision-making and accurate model interpretation. (1 pt)**

    1. Glimpse and Summary the Data (0.25 pt):

        - Use functions like `glimpse()` and `summary()` in R to get a quick overview of the dataset. This will provide information on variable types, missing values, and basic statistics.

```
setwd("your wd")
library(readxl)
library(tidyverse)
smartwatch <- read_excel("SmartWatch Data File.xlsx")
glimpse(smartwatch)
summary(smartwatch)
```

```
Rows: 1,000
Columns: 12
$ ConstCom  <dbl> 3, 6, 7, 7, 7, 2, 7, 7, 6, 5, 6, 7, 7, 7, 4, 6, 7, 5, 4, 4, ~
$ TimelyInf <dbl> 2, 6, 4, 5, 4, 1, 7, 5, 5, 6, 5, 1, 7, 5, 2, 3, 6, 4, 3, 5, ~
$ TaskMgm   <dbl> 3, 6, 4, 4, 2, 3, 3, 6, 6, 6, 6, 6, 4, 3, 3, 6, 3, 5, 4, 4, ~
$ DeviceSt  <dbl> 3, 6, 4, 5, 6, 2, 6, 5, 3, 3, 3, 4, 4, 5, 5, 4, 6, 1, 6, 1, ~
$ Wellness  <dbl> 2, 5, 6, 5, 3, 2, 3, 2, 5, 5, 7, 7, 4, 3, 6, 3, 4, 2, 5, 2, ~
$ Athlete   <dbl> 3, 3, 4, 4, 2, 4, 1, 2, 3, 5, 6, 6, 2, 2, 7, 3, 3, 1, 6, 4, ~
$ Style     <dbl> 3, 1, 1, 4, 4, 3, 4, 3, 5, 3, 6, 6, 4, 3, 6, 5, 3, 4, 6, 4, ~
$ AmznP     <dbl> 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, ~
$ Female    <dbl> 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, ~
$ Degree    <dbl> 1, 2, 1, 2, 1, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 2, 2, 1, 1, 1, ~
$ Income    <dbl> 2, 3, 3, 5, 3, 2, 5, 3, 2, 3, 3, 4, 3, 3, 2, 3, 3, 1, 2, 1, ~
$ Age       <dbl> 38, 38, 42, 35, 36, 47, 36, 41, 46, 28, 30, 32, 40, 37, 25, ~
```

```
    ConstCom       TimelyInf        TaskMgm         DeviceSt        Wellness    
 Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.000  
 1st Qu.:4.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.000  
 Median :5.00   Median :4.00   Median :4.00   Median :4.00   Median :5.000  
 Mean   :4.69   Mean   :4.29   Mean   :4.19   Mean   :3.84   Mean   :4.367  
 3rd Qu.:6.00   3rd Qu.:5.00   3rd Qu.:5.00   3rd Qu.:5.00   3rd Qu.:6.000  
 Max.   :7.00   Max.   :7.00   Max.   :7.00   Max.   :7.00   Max.   :7.000  
    Athlete          Style           AmznP           Female      
 Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :0.000  
 1st Qu.:3.000   1st Qu.:3.000   1st Qu.:0.000   1st Qu.:0.000  
 Median :4.000   Median :4.000   Median :1.000   Median :1.000  
 Mean   :3.814   Mean   :4.299   Mean   :0.564   Mean   :0.566  
 3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :7.000   Max.   :7.000   Max.   :1.000   Max.   :1.000  
     Degree          Income           Age       
 Min.   :1.000   Min.   :1.000   Min.   :24.00  
 1st Qu.:1.000   1st Qu.:3.000   1st Qu.:31.00  
 Median :1.000   Median :3.000   Median :36.00  
 Mean   :1.332   Mean   :3.292   Mean   :35.52  
 3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:40.00  
 Max.   :2.000   Max.   :5.000   Max.   :47.00  
```

2. Distinguish Bases Variables from Descriptive Variables (0.25 pt)

   - Bases variables include all variables except for `Female,Degree,Income,Age`

```
smartwatch_NonDemog = smartwatch %>% select(-c(Female,Degree,Income,Age))
```

3. Obtain Average of Bases Variables (0.25 pt):

- Calculate the average of the bases variables, grouping by income, female, and Amazon membership. This can be achieved using functions like `group_by()` and `summarize()` in R.

```
smartwatch %>% group_by(Income) %>% summarise(avg_ConstCom = mean(ConstCom),
                                              avg_TimelyInf = mean(TimelyInf),
                                              avg_TaskMgm = mean(TaskMgm),
                                              avg_DeviceSt = mean(DeviceSt),
                                              avg_Wellness = mean(Wellness),
                                              avg_Athlete = mean(Athlete),
                                              avg_Style = mean(Style))
```

```
# A tibble: 5 x 8
  Income avg_ConstCom avg_TimelyInf avg_TaskMgm avg_DeviceSt avg_Wellness
   <dbl>        <dbl>         <dbl>       <dbl>        <dbl>        <dbl>
1      1         4.15          3.62        3.38         2.88         3.48
2      2         4.11          3.93        3.96         3.18         4.02
3      3         4.73          4.30        4.09         3.74         4.23
4      4         4.67          4.30        4.36         3.79         4.57
5      5         5.32          4.77        4.66         4.96         5
# i 2 more variables: avg_Athlete <dbl>, avg_Style <dbl>
```

```
smartwatch %>% group_by(Female) %>% summarise(avg_ConstCom = mean(ConstCom),
                                              avg_TimelyInf = mean(TimelyInf),
                                              avg_TaskMgm = mean(TaskMgm),
                                              avg_DeviceSt = mean(DeviceSt),
                                              avg_Wellness = mean(Wellness),
                                              avg_Athlete = mean(Athlete),
                                              avg_Style = mean(Style))
```

```
# A tibble: 2 x 8
  Female avg_ConstCom avg_TimelyInf avg_TaskMgm avg_DeviceSt avg_Wellness
   <dbl>        <dbl>         <dbl>       <dbl>        <dbl>        <dbl>
1      0         4.59          4.20        4.14         3.80         4.15
2      1         4.77          4.36        4.23         3.87         4.53
# i 2 more variables: avg_Athlete <dbl>, avg_Style <dbl>
```

```
smartwatch %>% group_by(AmznP) %>% summarise(avg_ConstCom = mean(ConstCom),
                                             avg_TimelyInf = mean(TimelyInf),
                                             avg_TaskMgm = mean(TaskMgm),
                                             avg_DeviceSt = mean(DeviceSt),
```

```
                                                    avg_Wellness = mean(Wellness),
                                                    avg_Athlete = mean(Athlete),
                                                    avg_Style = mean(Style))
```

```
# A tibble: 2 x 8
  AmznP avg_ConstCom avg_TimelyInf avg_TaskMgm avg_DeviceSt avg_Wellness
  <dbl>        <dbl>         <dbl>       <dbl>        <dbl>        <dbl>
1     0         4.53          4.27        4.10         3.43         3.91
2     1         4.82          4.31        4.26         4.15         4.72
# i 2 more variables: avg_Athlete <dbl>, avg_Style <dbl>
```

4. Briefly Discuss Data Patterns (0.25 pt):
    - Analyze the obtained averages and look for patterns or trends. Consider how bases variables vary across different groups such as income levels, gender, and Amazon membership. Note any significant differences or similarities observed.

2. **Now, run the cluster analysis using the R function that we used in the class (3 pt)**
    1. Draw Elbow Plot with Bases Variables (0.5 pt):
        - Use the R function discussed in class draw an elbow plot using only the bases variables.
    2. Draw Elbow Plot with Base and Demographic Variables (0.5 pt):
        - Repeat the process, this time including both base and demographic variables.
    3. Determine Optimal Number of Clusters (0.5 pt):
        - Examine the elbow plot and determine the number of clusters where the rate of decrease sharply changes (elbow point). This is often considered the optimal number of clusters.
    4. Segment Size (0.25 pt):
        - Once you've determined the optimal number of clusters, check the size (number of consumers) in each segment.
    5. Create Creative Segment Names and Describe them (1 pt):
        - Assign creative and relevant names to each segment based on their unique characteristics. Describe the unique features or behaviors that define each cluster.
    6. Visualize Clusters (0.25 pt):

- Utilize the R function from our class to visualize the clusters. This can include creating scatter plots, dendrograms, or any other relevant visualization method to showcase the distinctiveness of each segment.

3. **Run the discriminant analysis using the same smartwatch data. Keep in mind that you should include only demographic variables (1 pt).**

    1. Utilize the R function discussed in class for this analysis (0.5 pt)
    2. Evaluate the significance of the discriminant function to ascertain its statistical importance (0.25 pt)
    3. Create a confusion table to scrutinize the hit rates. Given the use of the same dataset, anticipate achieving notably high accuracy rates. It's worth highlighting that during our class discussions, different dataset (specifically, prospect data) was employed for the segment prediction (0.25 pt)

4.