Estimating the Demand

A) The available dataset has 1000 sales records of previously sold items. Besides the percentage of sales in each hour and the total sold items, there are 2 categorical features which are department's name of the item and the part of the event's day. In this assignment, you should determine the demand for all available items in the dataset.

B) If the categorical features had more than 2 categories, would it be possible to use k-means clustering? Discuss your answer.

Answer A) From the 'Assignment6_Data.csv' dataset, there are 1,000 obs. of 29 variables. At first, we have imported the data and then converted the categorical data into numerical data.

```
> str(AS6Data)
'data.frame':   1000 obs. of  29 variables:
 $ ï..Item.              : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Department            : num  1 1 1 2 1 1 1 1 1 1 ...
 $ Event.Part.of.the.day: num  2 2 1 2 1 2 2 2 1 2 ...
 $ hour.1                : num  0.14 0.14 0.3 0.26 0.22 0.34 0.32 0.21 0.52 0.38 ...
 $ hour.2                : num  0.18 0.13 0.16 0.12 0.11 0.09 0.17 0.14 0.16 0.15 ...
 $ hour.3                : num  0.15 0.07 0.06 0.07 0 0.13 0.11 0.06 0 0.11 ...
 $ hour.4                : num  0.07 0.06 0.1 0.04 0.08 0.03 0.08 0.09 0.11 0.05 ...
 $ hour.5                : num  0 0.11 0.05 0.04 0 0.06 0.05 0.04 0.06 0.06 ...
 $ hour.6                : num  0 0.1 0.02 0.04 0.05 0.03 0.03 0.1 0 0.04 ...
 $ hour.7                : num  0.15 0.08 0.02 0.03 0 0.01 0.03 0.02 0 0.02 ...
 $ hour.8                : num  0.09 0 0.02 0.04 0.05 0.05 0.02 0.03 0.03 0 ...
 $ hour.9                : num  0.03 0 0.02 0.03 0 0.03 0.01 0.02 0.07 0.04 ...
 $ hour.10               : num  0.03 0.08 0.03 0 0.02 0.06 0.05 0.03 0.01 0.09 ...
 $ hour.11               : num  0.06 0 0.04 0.04 0.04 0.03 0.01 0.02 0 0 ...
 $ hour.12               : num  0.01 0 0.01 0.02 0 0.03 0.04 0.01 0 0 ...
 $ hour.13               : num  0.04 0.05 0.01 0.04 0.04 0.01 0.03 0 0 0.03 ...
 $ hour.14               : num  0.02 0.07 0.02 0.03 0 0.01 0 0.02 0 0 ...
 $ hour.15               : num  0 0 0 0.01 0.02 0.01 0 0 0 0 ...
 $ hour.16               : num  0 0 0 0 0.01 0 0.01 0 0 0 ...
 $ hour.17               : num  0 0 0 0 0 0.01 0.01 0.02 0 0 ...
 $ hour.18               : num  0 0 0 0 0 0 0 0 0 0.03 ...
 $ hour.19               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ hour.20               : num  0 0.01 0 0 0 0 0 0 0 0 ...
 $ hour.21               : num  0.01 0 0 0 0 0 0 0 0 0 ...
 $ hour.22               : num  0.01 0 0.02 0.01 0 0 0 0 0 0 ...
 $ hour.23               : num  0 0 0 0.01 0 0 0 0 0 0 ...
 $ hour.24               : num  0 0 0.02 0.02 0 0 0 0.03 0 0 ...
 $ Total.sales           : num  91 80 91 171 85 88 51 63 71 98 ...
 $ Total                 : num  91.8 80.8 91.6 171.6 85.4 ...
```

```
> summary(AS6Data)
    ï..Item.        Department    Event.Part.of.the.day     hour.1           hour.2           hour.3             hour.4
 Min.   :   1.0   Min.   :1.000   Min.   :1.000         Min.   :0.1100   Min.   :0.0400   Min.   :0.00000   Min.   :0.00000
 1st Qu.: 250.8   1st Qu.:1.000   1st Qu.:1.000         1st Qu.:0.2100   1st Qu.:0.1200   1st Qu.:0.05000   1st Qu.:0.05000
 Median : 500.5   Median :1.000   Median :2.000         Median :0.2800   Median :0.1400   Median :0.08000   Median :0.07000
 Mean   : 500.5   Mean   :1.396   Mean   :1.537         Mean   :0.2676   Mean   :0.1396   Mean   :0.07228   Mean   :0.07237
 3rd Qu.: 750.2   3rd Qu.:2.000   3rd Qu.:2.000         3rd Qu.:0.3100   3rd Qu.:0.1600   3rd Qu.:0.09000   3rd Qu.:0.08250
 Max.   :1000.0   Max.   :2.000   Max.   :2.000         Max.   :0.5200   Max.   :0.2200   Max.   :0.15000   Max.   :0.17000
     hour.5            hour.6            hour.7           hour.8            hour.9            hour.10            hour.11
 Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
 1st Qu.:0.04000   1st Qu.:0.02000   1st Qu.:0.0200   1st Qu.:0.02000   1st Qu.:0.02000   1st Qu.:0.01000   1st Qu.:0.01000
 Median :0.06000   Median :0.04000   Median :0.0300   Median :0.03000   Median :0.03000   Median :0.03000   Median :0.03000
 Mean   :0.05479   Mean   :0.04038   Mean   :0.0421   Mean   :0.04409   Mean   :0.02778   Mean   :0.02744   Mean   :0.02637
 3rd Qu.:0.07000   3rd Qu.:0.05000   3rd Qu.:0.0500   3rd Qu.:0.06000   3rd Qu.:0.03000   3rd Qu.:0.03000   3rd Qu.:0.04000
 Max.   :0.13000   Max.   :0.12000   Max.   :0.1500   Max.   :0.23000   Max.   :0.07000   Max.   :0.09000   Max.   :0.06000
    hour.12           hour.13           hour.14           hour.15           hour.16           hour.17           hour.18
 Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
 1st Qu.:0.01000   1st Qu.:0.01000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
 Median :0.03000   Median :0.03000   Median :0.01000   Median :0.01000   Median :0.01000   Median :0.00000   Median :0.00000
 Mean   :0.02186   Mean   :0.02113   Mean   :0.01375   Mean   :0.00959   Mean   :0.01023   Mean   :0.00605   Mean   :0.00446
 3rd Qu.:0.04000   3rd Qu.:0.03000   3rd Qu.:0.02000   3rd Qu.:0.01000   3rd Qu.:0.01000   3rd Qu.:0.01000   3rd Qu.:0.00000
 Max.   :0.07000   Max.   :0.05000   Max.   :0.07000   Max.   :0.03000   Max.   :0.09000   Max.   :0.05000   Max.   :0.05000
    hour.19           hour.20           hour.21           hour.22           hour.23           hour.24          Total.sales
 Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   : 51.00
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.: 64.00
 Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000   Median : 79.00
 Mean   :0.00466   Mean   :0.00513   Mean   :0.00862   Mean   :0.00679   Mean   :0.00324   Mean   :0.00495   Mean   : 89.62
 3rd Qu.:0.00000   3rd Qu.:0.01000   3rd Qu.:0.01000   3rd Qu.:0.01000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:103.00
 Max.   :0.06000   Max.   :0.08000   Max.   :0.05000   Max.   :0.05000   Max.   :0.03000   Max.   :0.05000   Max.   :488.00
     Total
 Min.   : 51.63
 1st Qu.: 64.63
 Median : 79.59
 Mean   : 90.29
 3rd Qu.:103.67
 Max.   :488.72
```

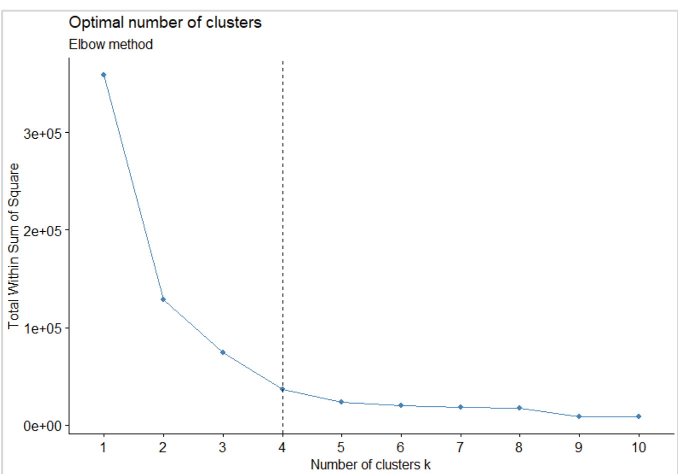
Here are the 2 categorical features which are department's name of the item and the part of the event's day. Please refer to the R code as determined the demand for all available items in the dataset.

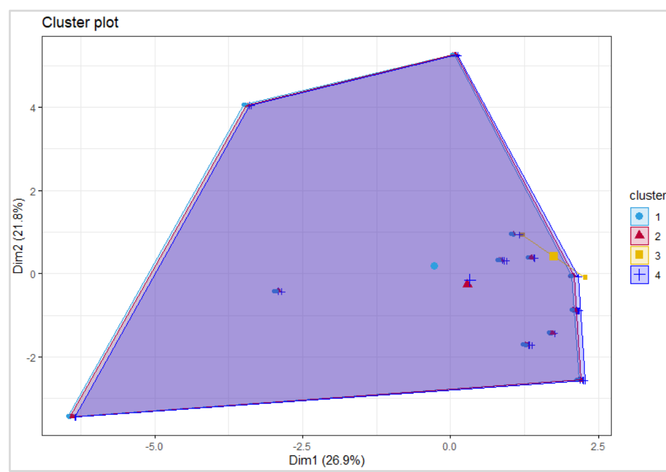
```
true_demand
stockouts=cbind(stockouts,true_demand)
View(stockouts)
colnames(stockouts)
colnames(not_stockouts)
View(stockouts[,c(2,31)])
```

Output:

```
> colnames(stockouts)
 [1] "ï..Item."          "Department"        "Event.Part.of.the.day" "hour.1"         "hour.2"
 [6] "hour.3"            "hour.4"            "hour.5"            "hour.6"            "hour.7"
[11] "hour.8"            "hour.9"            "hour.10"           "hour.11"           "hour.12"
[16] "hour.13"           "hour.14"           "hour.15"           "hour.16"           "hour.17"
[21] "hour.18"           "hour.19"           "hour.20"           "hour.21"           "hour.22"
[26] "hour.23"           "hour.24"           "Total.sales"       "Total"             "stockout_time"
[31] "true_demand"
```

```
> #K-means Clustering
> clusters <- kmeans(not_stockouts[,3:28],4, nstart = 20)
> not_stockouts=cbind(not_stockouts,clusters$cluster)
> centroids=clusters$center
> centroids[4,]
Event.Part.of.the.day          hour.1          hour.2          hour.3          hour.4
         1.677419e+00    2.664516e-01    1.441935e-01    7.290323e-02    6.032258e-02
               hour.5          hour.6          hour.7          hour.8          hour.9
         4.677419e-02    3.838710e-02    4.032258e-02    3.000000e-02    2.258065e-02
              hour.10         hour.11         hour.12         hour.13         hour.14
         2.290323e-02    2.709677e-02    2.064516e-02    2.225806e-02    1.290323e-02
              hour.15         hour.16         hour.17         hour.18         hour.19
         7.419355e-03    1.032258e-02    5.161290e-03    5.806452e-03    3.870968e-03
              hour.20         hour.21         hour.22         hour.23         hour.24
         4.193548e-03    8.387097e-03    7.419355e-03    6.774194e-03    2.322581e-02
          Total.sales
         1.614839e+02
```





Answer B) It is simply not possible to use the k-means clustering over categorical data because we need a distance between elements and that is not clear with categorical data as it is with the numerical part of our data. So, the best solution that comes to my mind is that we construct somehow a similarity matrix (or dissimilarity/ distance matrix) between our categories to complement it with the distances for our numerical data (for which we can use simply an euclidean or manhattan distance). Then use the K-medoid algorithm, which can accept a dissimilarity matrix as input and using R with the "cluster" package that includes the pam() function.

If there is a logical order of the categories (i.e. category A is more similar to category B than to category C due to some features of the categories) we can apply weighted values to categories. But this is a typical "false" category feature (because it can be decomposed in a vector of numerical features). If the problem is related to real categorical features each category has the same distance to each other. You can set a fixed distance for any category feature depending on the logic importance (weight) of this category for clustering.

Example: if you have two category features A and B by the knowledge of the clustering problem, we can set that a mismatch in the category.

Bibliographic Reference:

(Anonymous, 2021) *Colors in R* Retrieved from http://www.sthda.com/english/wiki/colors-in-r

Shendre S. April 29, 2020 *Clustering datasets having both numerical and categorical variables* Retrieved from https://towardsdatascience.com/clustering-datasets-having-both-numerical-and-categorical-variables-ed91cdca0677

Marchese L. and Ramirez-Flandes S. July 21, 2021 *RESEARCH-GATE* Retrieved from https://www.researchgate.net/post/What-is-the-best-way-for-cluster-analysis-when-you-have-mixed-type-of-data-categorical-and-scale