Untitled

Ray Kwon

2023-01-09

1 Introduction

These are two datasets with Portuguese hotel demand data and each observation represents a hotel booking. One of the hotels is a resort hotel(40,060 observations) and other is a city hotel(79,330 observations).

1.1 Problem Statement :

What differentiated marketing strategy can attract the exiting customers ?

1.2 Goal:

Segmenting all customers using K-mean clustering and then identifying recommendations for hotels by analyzing customer needs.

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

```
#data minig

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
```

```r
hotel <- read.csv("hotel_bookings.csv")


resort <- filter(hotel, hotel == "Resort Hotel")

resort$meal_num <-  ifelse(resort$meal == 'BB',1,ifelse(resort$meal == 'FB',2,ifelse(resort$meal == "HB

resort$depo_num <- ifelse(resort$deposit_type == "No Deposit", 1, ifelse(resort$deposit_type == "Refunda


resort$family <- ifelse(resort$children > 0 | resort$babies > 0, 1,0)


resort_demand <- filter(resort, adr > 20)
```

1.3 Normalization

Distance of clusters is highly influenced by scale of variables, it is customary to normalize first. In our dataset, all variables are on not same scale(0,1), so normalization is necessary:

```
## Loading required package: ggplot2

## Loading required package: lattice
```

2 Exploratory Data Analysis (EDA)

There are 5 variables to figure out marketing strategy: Family & Meal Plan & Type of deposit & ADR & Paring space

Family: the customers brought their kids ?

*Accompanied kids: 1 & NO accompanied kids: 0

Meal Plan: what type of meal plan the customers prefered?

*No plan: 0 & BB(Breakfast and Bed): 1 & FB(Full Board):2 & HB(Half Board): 3 & SC(Self Catering): 4

Type of Deposit: What type of deposit the customers chose ?

- No deposit: 1 & Refundable: 2 & Non Refund:3

Parking: How many parking spaces were required by customers?

ADR: Average Daily Rate(Dividing room revenue by rooms sold)

2.1 How many kids are accompanied by customers? :

About 90% of customers didn't bring children or babies and few customers brought 1 or 2 children

2.2 Types of deposit

About 90% of customers chose 1(No Deposit)

2.3 Types of Meal plan

About 70% of customers chose 1(Breakfast and beds) and about 20% of customer chose 3(Half Board) that means including two meals

2.4 Average Daily Rate

Each room generates $50 to $100 per a day mostly

3 Elbow Curve

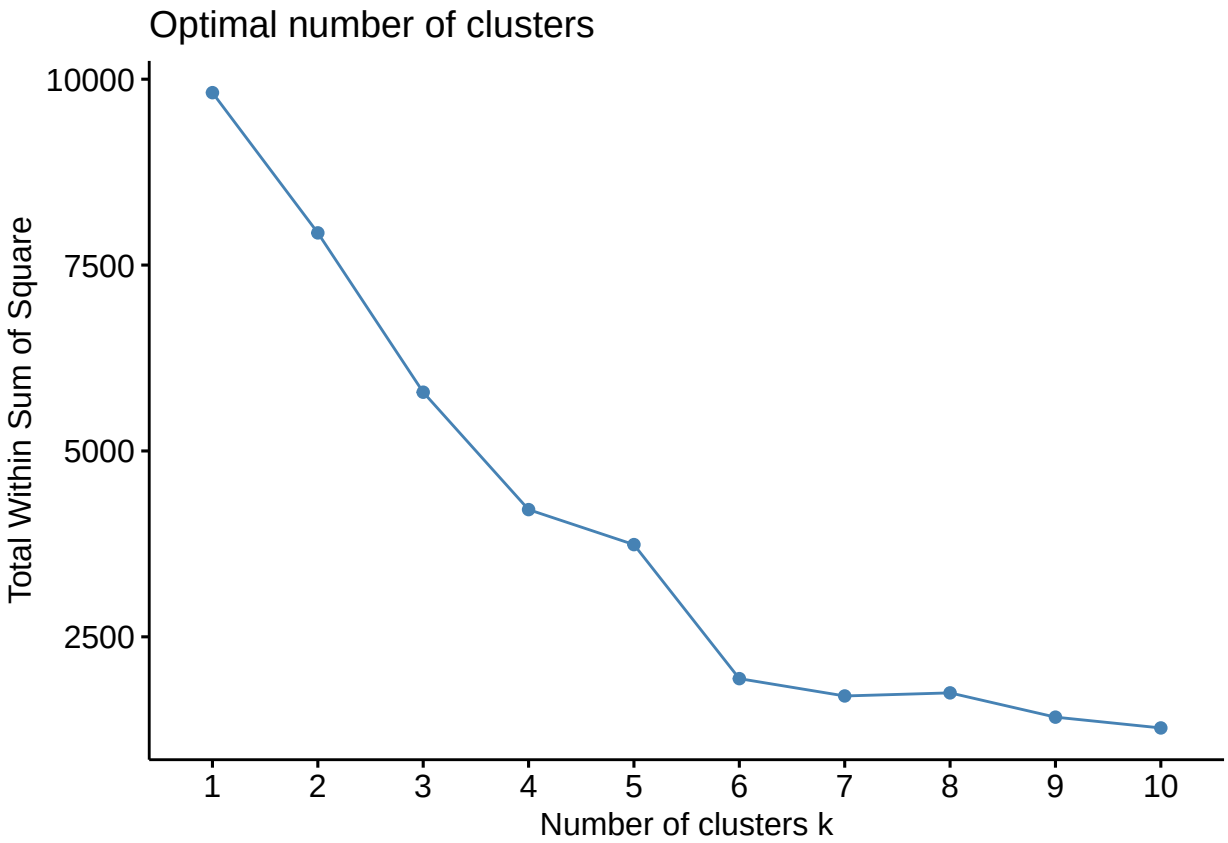
3.1 why we need to use elbow curve ?

We need use elbow curve to figure out what is the optimal number of cluster. The graph looks like elbow shape and we need to find elbow point that is k-value where the elbow gets created. This is because increasing the value of "K" does not reduce WCSS(Within-Cluster Sum of Square)

3.2 Optimal number of cluster

We choose the elbow point which is 6. Even though we increased the k-value, it does not reduce WCSS

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```



4 Result of K-mean Clustering

4.1 Size of Clusters

Cluster1: 3,868 Cluster2: 4,627 Cluster3: 5,413 Cluster4: 17,487 Cluster5: 5,892 Cluster6: 1,892

4.1 Insight

°Cluster1 & Cluster3 show the highest average daily rate

°All clusters chose "No deposit" except cluster 6 that chose "Non refund" option

°Customers in cluster1 brought their kids

°Most customers in cluster5 chose Half Board, whereas, most customers in rest clusters chose "Breakfast and Bed" option

°Cluster2 required parking space

```
##       family   meal_num   depo_num         adr required_car_parking_spaces
## 1  3.0147988  0.1397502 -0.2196305  1.09393220                   0.1735987
## 2 -0.3280724 -0.1281032 -0.2119857  0.08157423                   2.4684493
## 3 -0.3316886 -0.4902237 -0.2190932  1.10299525                  -0.3940615
## 4 -0.3316886 -0.5107956 -0.2208826 -0.62542327                  -0.3951103
## 5 -0.3316886  1.8924213 -0.2200606  0.21926567                  -0.3951103
## 6 -0.3243699  0.2667564  4.4699351 -0.51089310                  -0.3951103
```

```
##       family   meal_num   depo_num         adr required_car_parking_spaces
## 1  3.0147988  0.1397502 -0.2196305  1.09393220                   0.1735987
## 2 -0.3280724 -0.1281032 -0.2119857  0.08157423                   2.4684493
## 3 -0.3316886 -0.4902237 -0.2190932  1.10299525                  -0.3940615
## 4 -0.3316886 -0.5107956 -0.2208826 -0.62542327                  -0.3951103
## 5 -0.3316886  1.8924213 -0.2200606  0.21926567                  -0.3951103
## 6 -0.3243699  0.2667564  4.4699351 -0.51089310                  -0.3951103
```

```
##
##     1     2     3     4     5     6
##  3868  4627  5413 17487  5892  1829
```

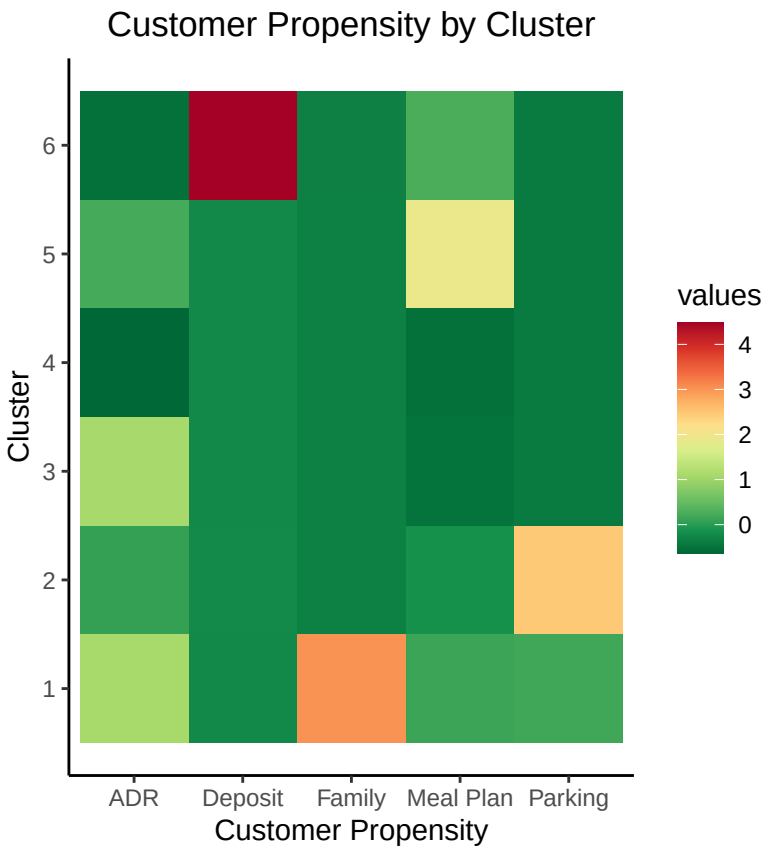
```
##           1           2           3           4           5           6
## 1.000000000 0.001080614 0.000000000 0.000000000 0.000000000 0.002186987
```

```
##         1         2         3         4         5         6
## 163.30238 102.13273 163.85000  59.41386 110.45245  66.33411
```

```
##        1        2        3        4        5        6
## 1.000517 1.003674 1.000739 1.000000 1.000339 2.937124
```

```
##         1         2         3         4         5         6
## 1.5144778 1.2896045 0.9855902 0.9683193 2.9859131 1.6211044
```

```
##            1            2            3            4            5            6
## 0.2003619442 1.0088610331 0.0003694809 0.0000000000 0.0000000000 0.0000000000
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.