7 . HIERARCHIAL CLUSTERING

23CSEG28

```
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

library(ggplot2)
library(ggcorrplot)
library(reshape2)
df=read.csv("C:/Users/HP/Downloads/fulfilment_center_info.csv")
head(df)

##   center_id city_code region_code center_type op_area
## 1        11       679          56      TYPE_A     3.7
## 2        13       590          56      TYPE_B     6.7
## 3       124       590          56      TYPE_C     4.0
## 4        66       648          34      TYPE_A     4.1
## 5        94       632          34      TYPE_C     3.6
## 6        64       553          77      TYPE_A     4.4

str(df)

## 'data.frame':    77 obs. of  5 variables:
##  $ center_id  : int  11 13 124 66 94 64 129 139 88 143 ...
##  $ city_code  : int  679 590 590 648 632 553 593 693 526 562 ...
##  $ region_code: int  56 56 56 34 34 77 77 34 34 77 ...
##  $ center_type: chr  "TYPE_A" "TYPE_B" "TYPE_C" "TYPE_A" ...
##  $ op_area    : num  3.7 6.7 4 4.1 3.6 4.4 3.9 2.8 4.1 3.8 ...

df$center_type=as.factor(df$center_type)
df$center_id=as.factor(df$center_id)
df$city_code=as.factor(df$city_code)
df$region_code=as.factor(df$region_code)
summary(df)

##    center_id     city_code    region_code center_type    op_area     
##  10     : 1   590    : 9   56     :30   TYPE_A:43   Min.   :0.900  
##  11     : 1   526    : 8   34     :21   TYPE_B:15   1st Qu.:3.500  
##  13     : 1   638    : 3   77     :17   TYPE_C:19   Median :3.900  
##  14     : 1   517    : 2   85     : 5               Mean   :3.986  
##  17     : 1   522    : 2   23     : 1               3rd Qu.:4.400  
##  20     : 1   576    : 2   35     : 1               Max.   :7.000
##  (Other):71   (Other):51   (Other): 2
```

```
colSums(is.na(df))
```

```
##   center_id   city_code region_code center_type     op_area
##           0           0           0           0           0
```

```
# univariate analysis
ggplot(df,aes(center_type))+geom_bar()+
  labs(title="Distribution of center_type")
```

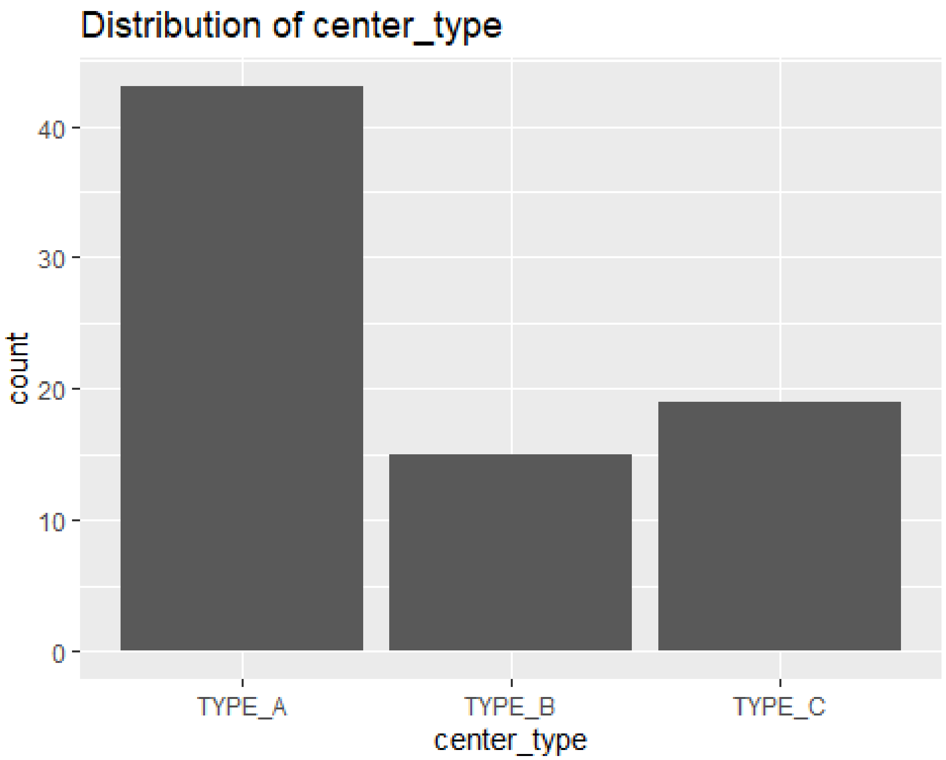

Fig 7.1

```
ggplot(df,aes(x=op_area))+geom_histogram(binwidth=1)+
  labs(title ="Distribution of Operational Area",xlab="operational_area",col=
'yellow')
```

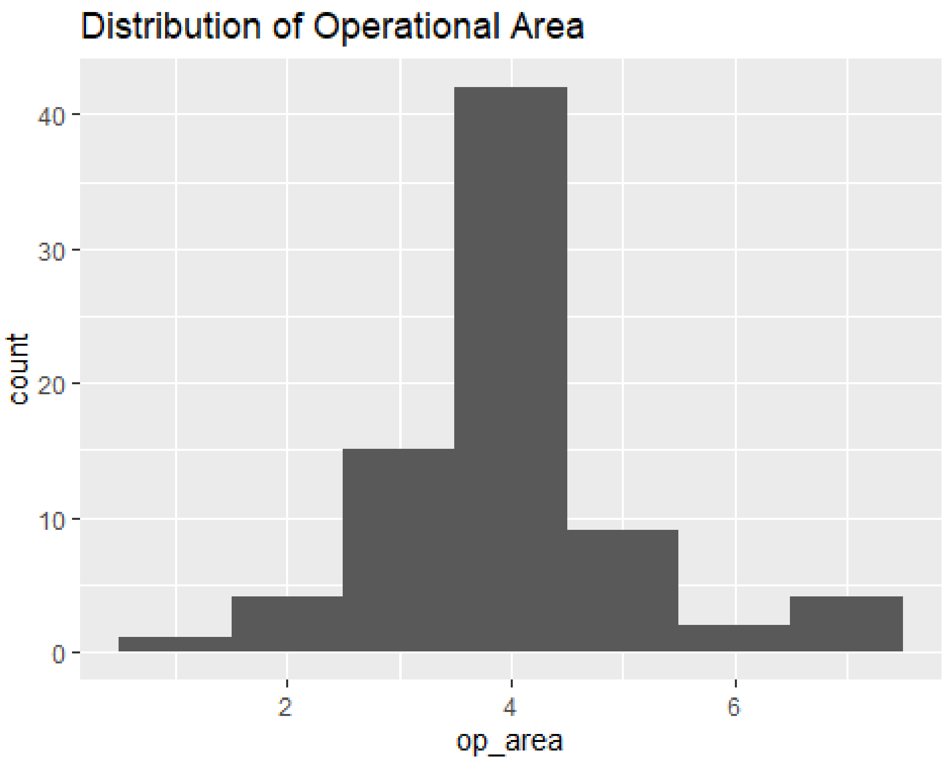

Fig 7.2

```r
#Bivariate Analaysis
ggplot(df,aes(x=center_type,y=op_area))+geom_bar(stat = "identity")+
  labs(title="centre type vs operation area")
```

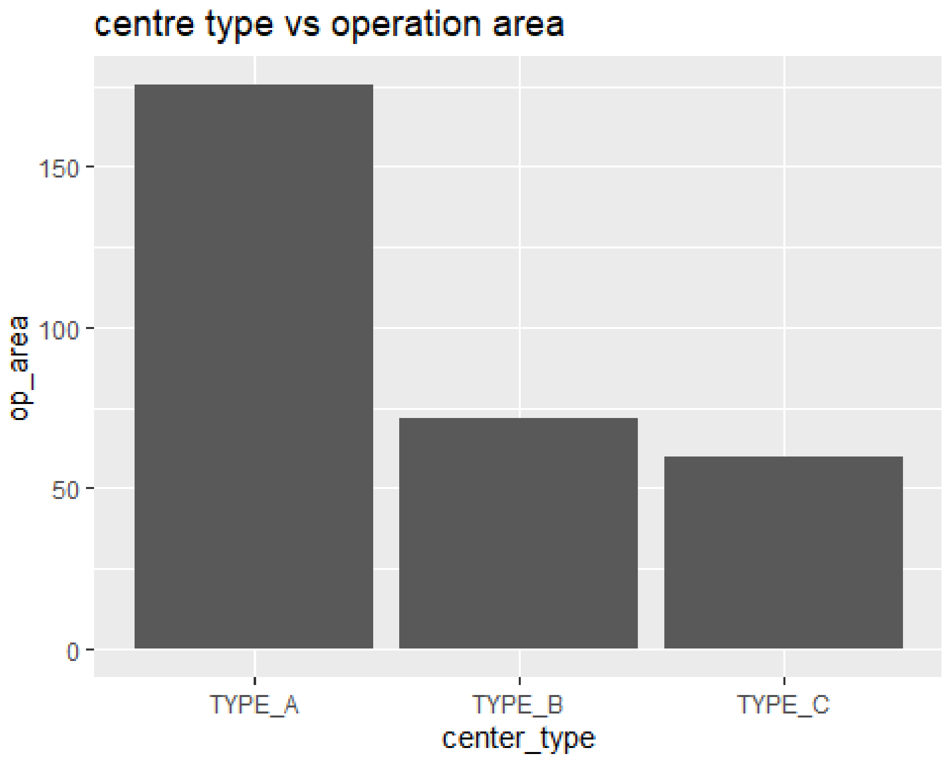

Fig 7.3

```r
df%>%
  group_by(center_id)%>%
  summarise(n=mean(op_area))%>%
  filter(n>5)%>%
  ggplot(aes(x=reorder(center_id,-n),y=n))+geom_bar(stat="identity")+labs(tit
le="centre id vs operation area")
```

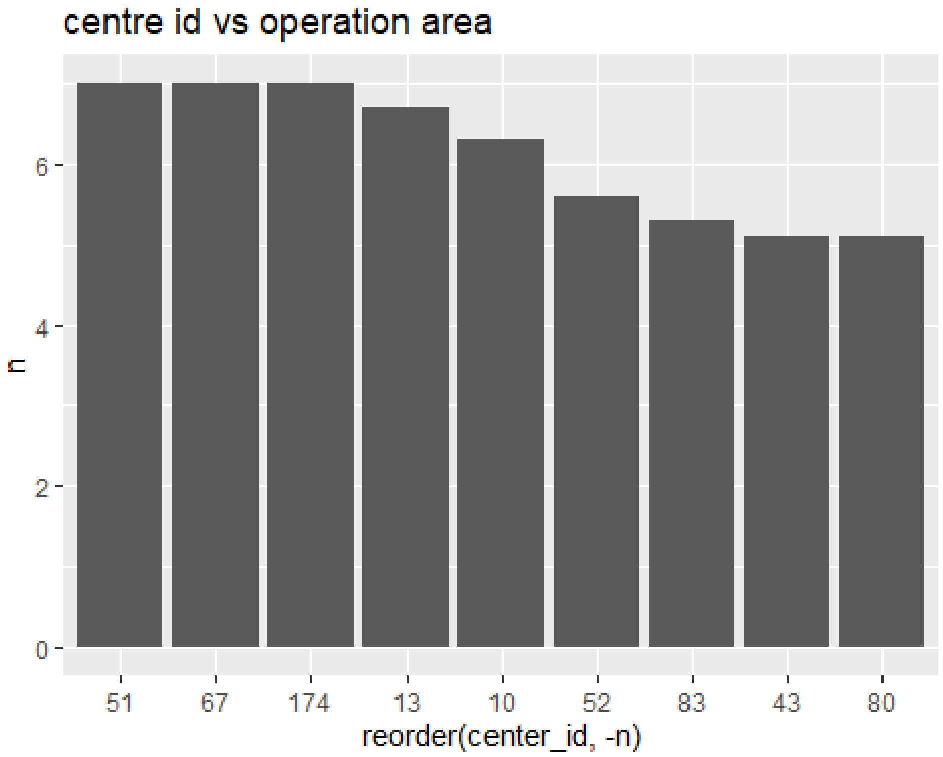

Fig 7.4

```r
df%>%
  group_by(region_code)%>%
  summarise(n=mean(op_area))%>%
  ggplot(aes(x=reorder(region_code,-n),y=n))+geom_bar(stat="identity")+
  labs(title="region code vs operation area")
```

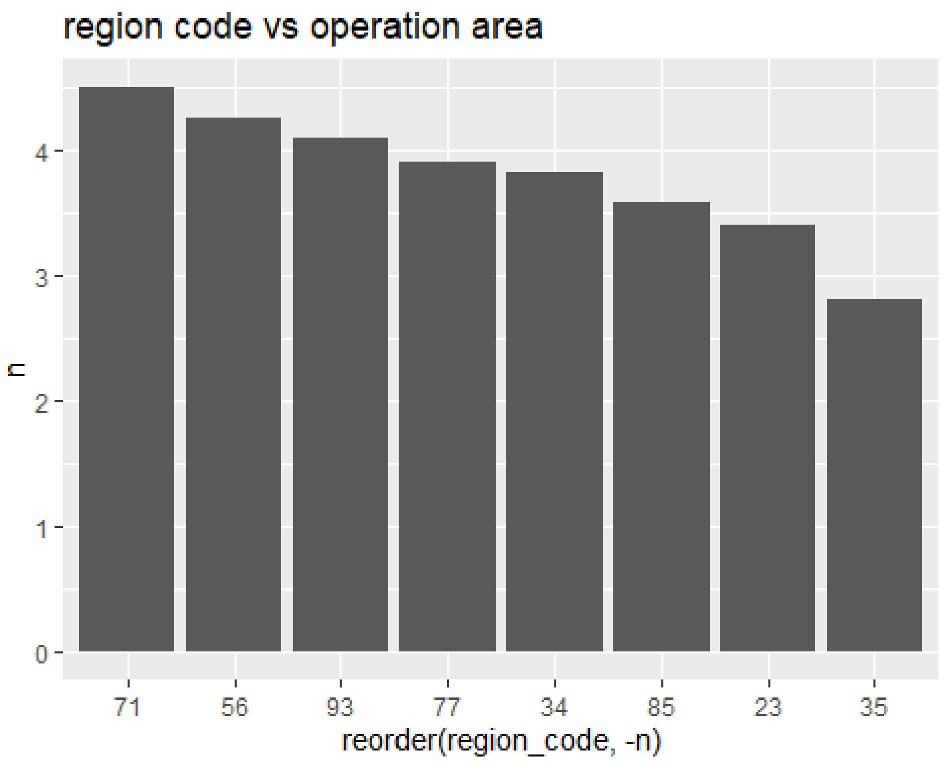

Fig 7.5

```
df%>%
  group_by(city_code)%>%
  summarise(n=mean(op_area))%>%
  filter(n>4)%>%
  ggplot(aes(x=reorder(city_code,-n),y=n))+geom_bar(stat="identity")+
  labs(title="city_code vs op_area")
```

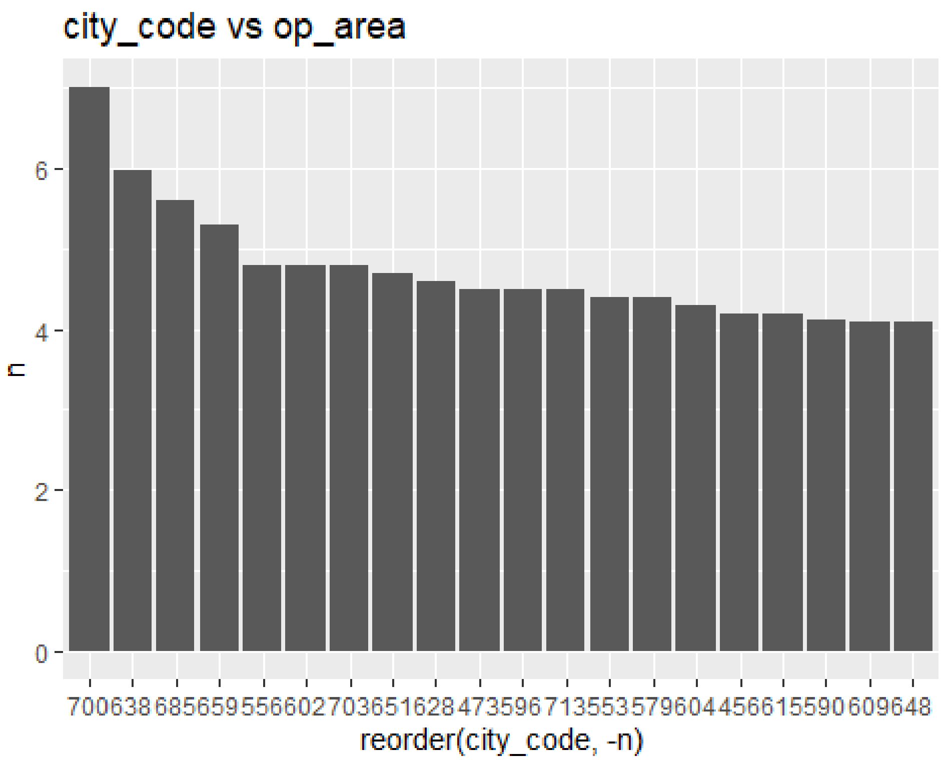

Fig 7.6

```
df[sapply(df, is.factor)] <- data.matrix(df[sapply(df, is.factor)])
data=cor(df[sapply(df, is.numeric)])
data1= melt(data)
ggcorrplot(data, hc.order = TRUE,lab = TRUE)
```

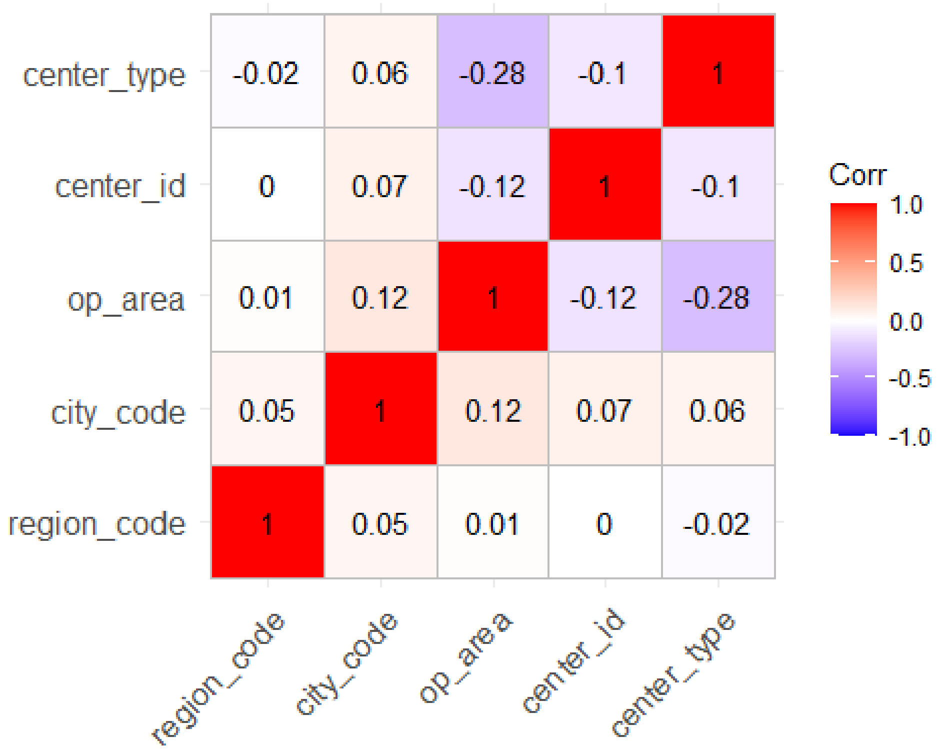

Fig 7.7

```
#Hierarchical clustering
dist_met=dist(df,method="euclidean")
set.seed(50)
clust=hclust(dist_met,method='ward.D2')
clust

##
## Call:
## hclust(d = dist_met, method = "ward.D2")
##
## Cluster method   : ward.D2
## Distance         : euclidean
## Number of objects: 77

plot(clust,hang=-0.4,main="heirarchical clustering")
```

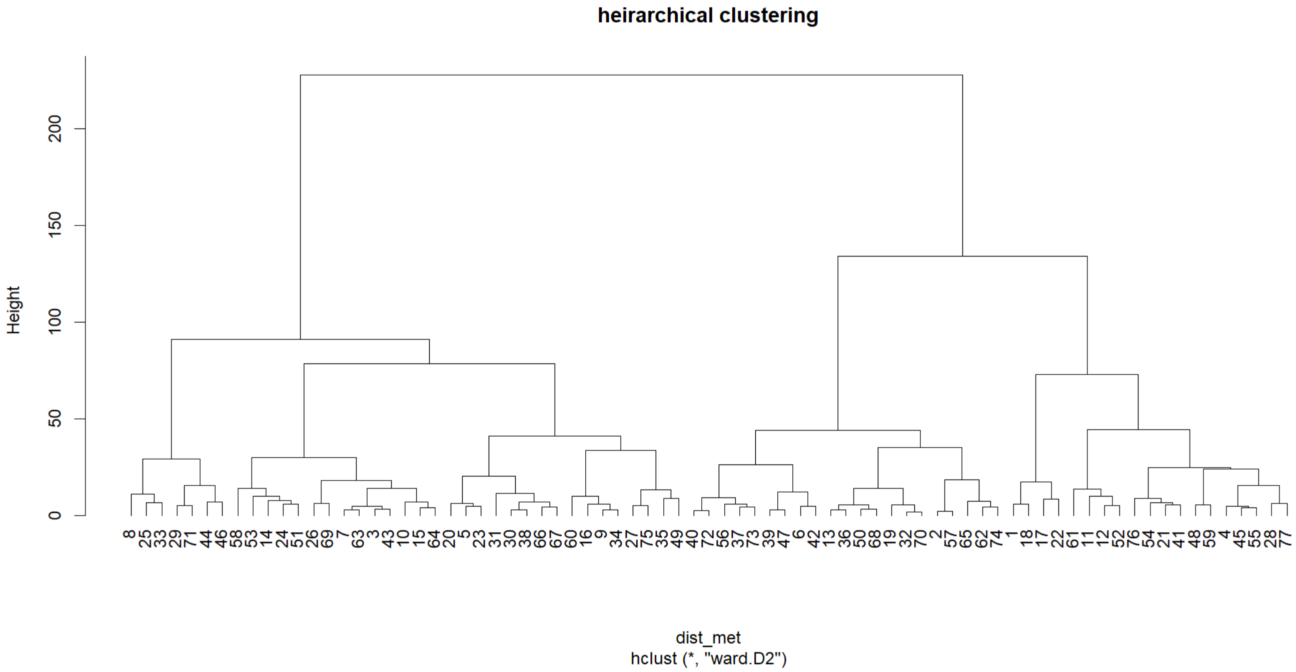
heirarchical clustering
Height
0
50
100
150
200
8
25
33
29
71
44
46
58
53
14
24
51
26
69
7
63
3
43
10
15
64
20
5
23
31
30
38
66
67
60
16
9
34
27
75
35
49
40
72
56
37
73
39
47
6
42
13
36
50
68
19
32
70
2
57
65
62
74
1
18
17
22
61
11
12
52
76
54
21
41
48
59
4
45
55
28
77
dist_met
hclust (*, "ward.D2")

Fig 7.8