3주차 패키지 오답노트

Part1. Anomaly detection

문제 0. (기본세팅) 워킹디렉토리를 설정하고 tidyverse, data.table, magrittr 패키지를 로드한 뒤 “part1_data.csv”데이터를 불러주세요.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(data.table)
```

```
##
## 다음의 패키지를 부착합니다: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(magrittr)
```

```
##
## 다음의 패키지를 부착합니다: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
data<-fread("part1_data.csv")
```

문제 1. 데이터를 자유롭게 파악하고, NA 가 있는지 확인해보세요.

```
summary(data)
```

```
##      sales           oil_price       onpromotion
##  Min.   :  11914   Min.   : 26.19   Min.   :   12
##  1st Qu.: 562280   1st Qu.: 44.71   1st Qu.: 1885
##  Median : 651852   Median : 48.61   Median : 5036
##  Mean   : 673832   Mean   : 53.08   Mean   : 5900
##  3rd Qu.: 782642   3rd Qu.: 53.76   3rd Qu.: 8862
##  Max.   :1285882   Max.   :106.06   Max.   :23403
##  NA's   :84        NA's   :19       NA's   :65
```

```
str(data)
```

```
## Classes 'data.table' and 'data.frame':   1139 obs. of  3 variables:
##  $ sales      : num  690270 719638 560254 574178 783741 ...
##  $ oil_price  : num  106 105 105 105 105 ...
##  $ onpromotion: int  2171 4270 2224 4017 569 540 460 2029 4247 2323 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
colSums(is.na(data))
```

```
##       sales   oil_price onpromotion
##          84          19          65
```

문제2. VIM 패키지를 활용하여 다음과 같이 NA 의 패턴을 시각화하고 해석해주세요.

```
library(VIM)
```

```
## Warning: 패키지 'VIM'는 R 버전 4.1.3에서 작성되었습니다
```

```
## 필요한 패키지를 로딩중입니다: colorspace
```

```
## 필요한 패키지를 로딩중입니다: grid
```

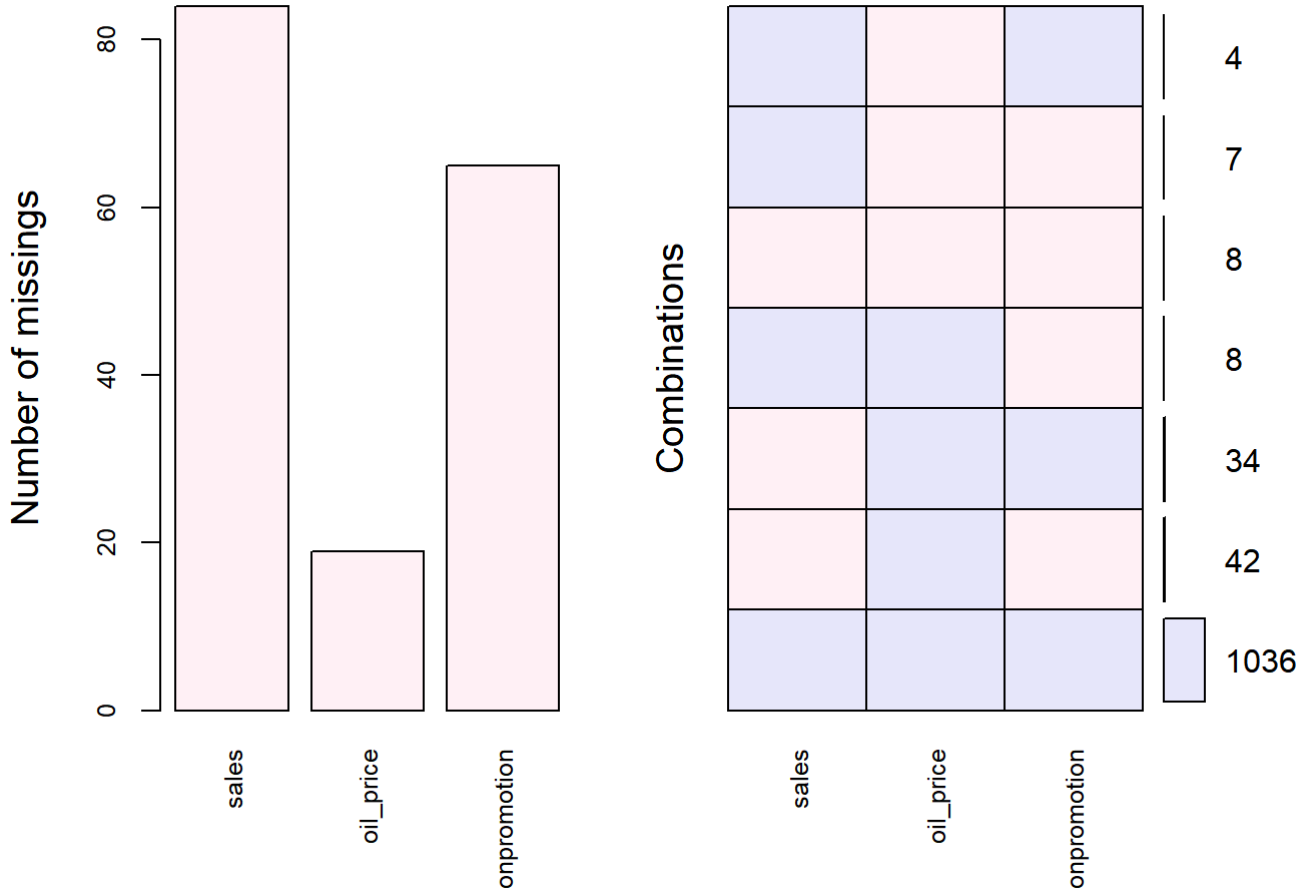
```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## 다음의 패키지를 부착합니다: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
aggr(data,prop=F,numbers=T,col=c('lavender','lavenderblush'),cex.axis=0.77)
```

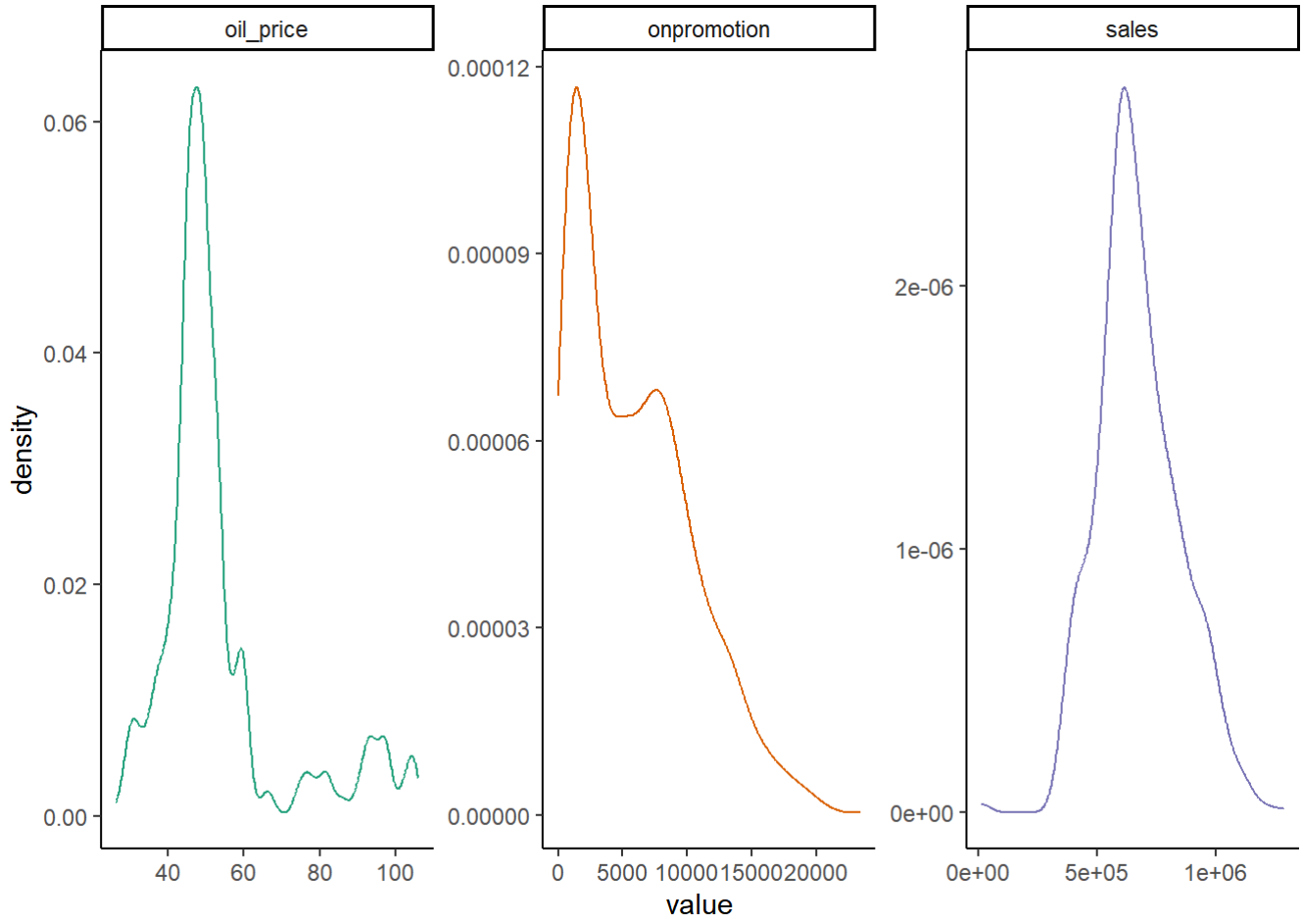


앞서 colSums 함수를 통해 파악한 결측치와 동일하게 그래프가 그려졌다. 두번째 플롯은 각 행에서 NA가 나타나는 패턴 별 수를 나타낸 플랏이다. 플랏에서 빨강색은 해당 값이 NA라는 것을 나타내고, 파랑색은 NA가 아니라는 것을 나타낸다. 확인해본 결과, sales와 onpromotion 데이터가 모두 NA인 것이 42개로 가장 많았고, 그 다음으로는 sales만 NA인 것은 34개로 많이 나왔고, 다른 패턴은 모두 10개 이하로 데이터에 NA를 포함하고 있었다. 이를 통하여, sales와 onpromotion의 NA가 많은 이유를 2개의 변수에서 동시에 NA가 발생하는 데이터가 많기 때문임을 확인할 수 있었다.

문제 3. 세 변수의 분포를 살펴보겠습니다. 아래와 같은 density plot 을 그린 후 간단히 해석해보세요.

```
data %>% gather %>% ggplot(aes(x=value,color=key))+
  geom_line(stat='density')+
  theme_classic()+
  scale_color_brewer(palette="Dark2")+
  theme(legend.position = "none")+
  facet_wrap(vars(key),ncol=3,scales="free")
```

```
## Warning: Removed 168 rows containing non-finite values (stat_density).
```

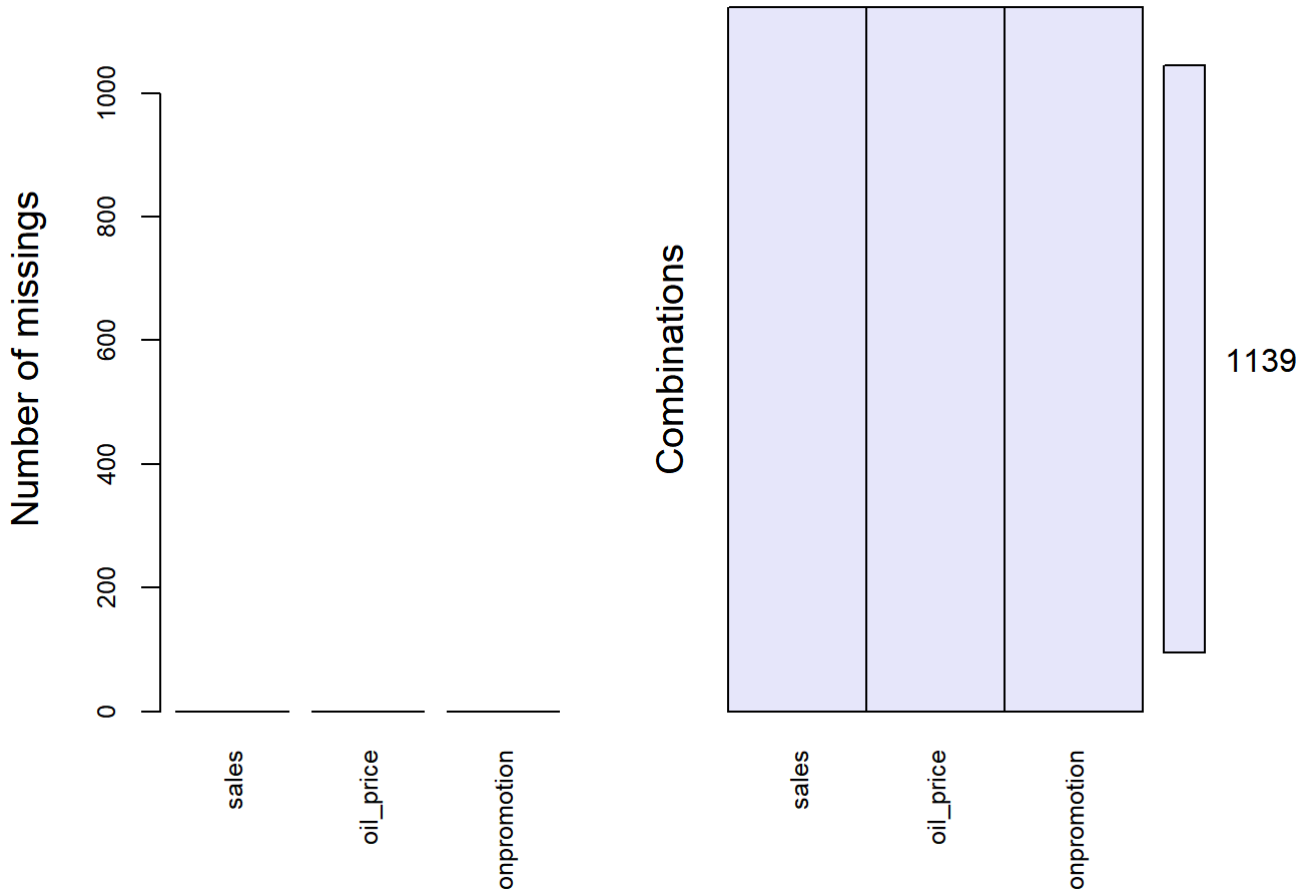


문제 4. 세 변수에 mean imputation 을 해준 뒤 다시 아래의 플랏을 다시 그려보세요.

```
meanin<-function(x){
  imp<-ifelse(is.na(x),mean(x,na.rm=T),x)
  return(imp)
}

data<-apply(data,2,meanin) %>% as.data.frame
```

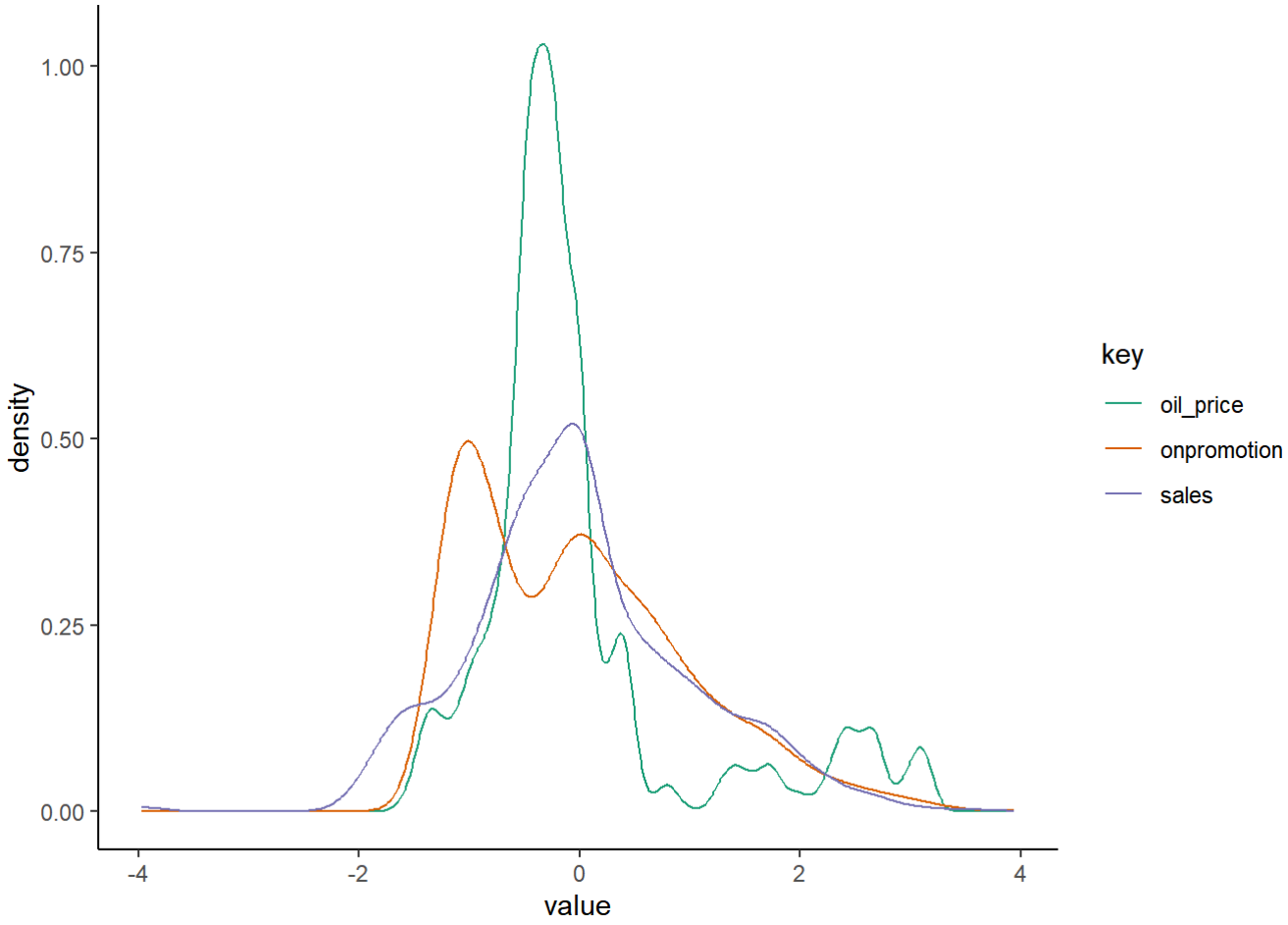
```
aggr(data,prop=F,numbers=T,col=c('lavender','lavenderblush'),cex.axis=0.77)
```



문제 5. 세 변수를 모두 scale 함수로 표준화 해준 뒤 아래의 플랏을 그려 분포를 확인해보세요.

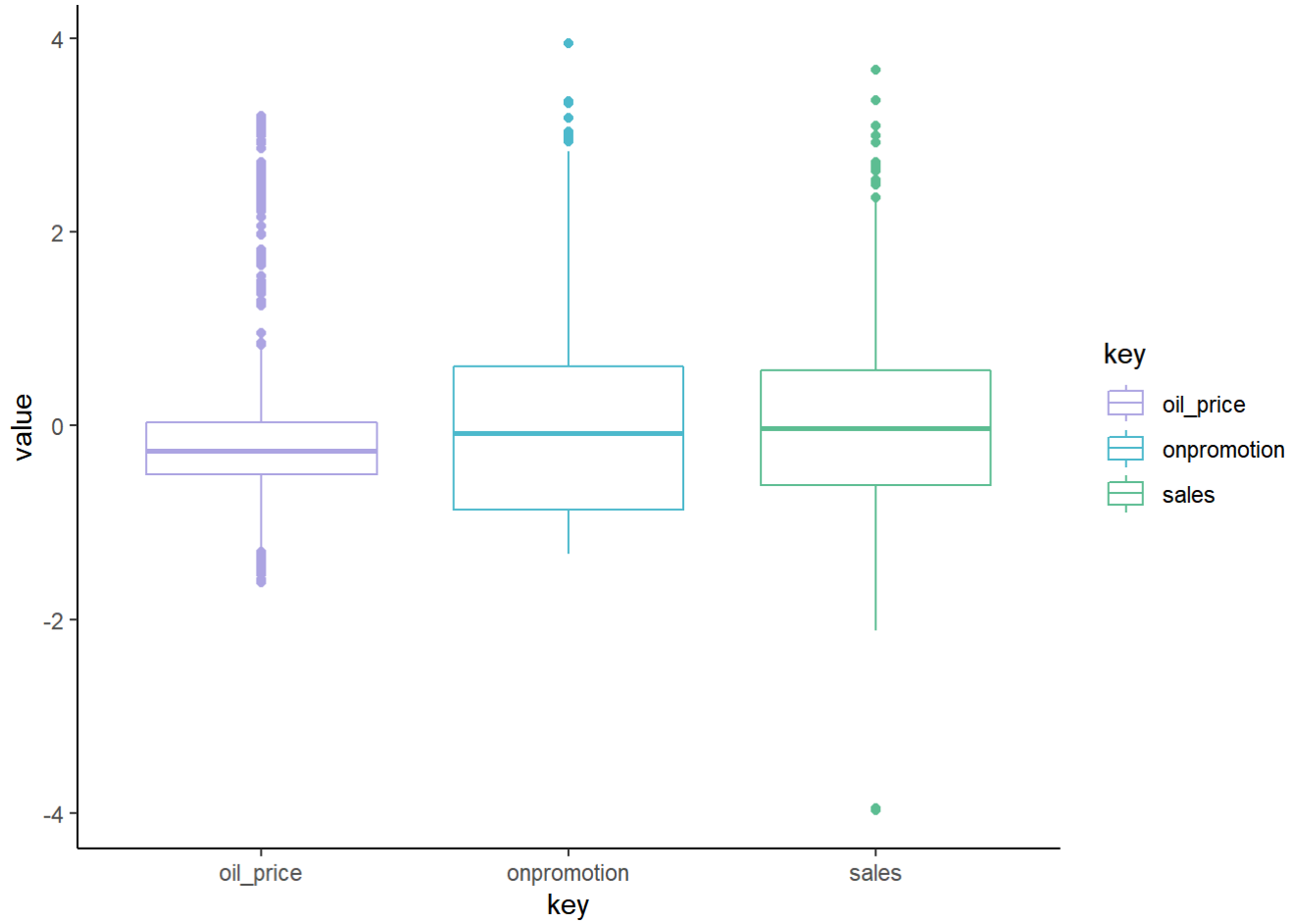
```
data %<>% scale %>% as.data.frame

data %>% gather %>%
  ggplot(aes(x=value,color=key))+
  geom_line(stat='density')+
  theme_classic()+
  scale_color_brewer(palette="Dark2")
```



문제 6. 아래의 boxplot 을 그려보고 각 변수의 이상치들을 한번 살펴보겠습니다.

```
col <- hcl.colors(3, palette ="Cold")
data %>% gather %>%
  mutate(key=as.factor(key)) %>%
  ggplot(aes(key,value,color=key))+
  geom_boxplot()+
  theme_classic()+
  scale_color_manual(values = col)
```



문제 7. train set 과 test set 을 분리하겠습니다. - rsample 패키지의 initial_split 사용 (prop = 0.7) - initial_split 과 training(), testing()을 사용해 각각 train, test 라는 데이터 프레임을 만들어주세요.

```
set.seed(2829)
library(rsample)
```

```
## Warning: 패키지 'rsample'는 R 버전 4.1.3에서 작성되었습니다
```

```
data_split<-initial_split(data,prop=0.7)
train<-training(data_split)
test<-testing(data_split)
```

문제 8. Isolation Forest

0. Isolation forest 에 대해 간단하게 알아보고 대표적인 파라미터에는 어떤 것들이 있는지 간단히 적어보세요.(이 문제는 선택입니다. 그러나 알아보고 공부하는건 꼭 개인적으로 하세요!)

Isolation Forest는 Unsupervised Anomaly Detection 중 하나로 현재 갖고 있는 데이터 중 이상치를 탐지할 때 주로 사용됩니다. 이름에서 볼 수 있듯이 tree 기반으로 구현되는데, 랜덤으로 데이터를 split하여 모든 관측치를 고립시키며 구현됨

변수가 많은 데이터에서도 효율적으로 작동할 수 있는 장점이 있음

컨셉 : “각 관측치를 고립(=분리)시키기는 것은 이상치가 정상 데이터보다 쉽다.”

파라미터 : https://hongl.tistory.com/150 (https://hongl.tistory.com/150)

1. isotree 패키지를 불러오세요.

```
library(isotree)
```

```
## Warning: 패키지 'isotree'는 R 버전 4.1.3에서 작성되었습니다
```

2. train 데이터를 활용해 isolation.forest 함수로 모델을 만들어보세요.(sample_size=64)

```
set.seed(2829)
model<-isolation.forest(train,sample_size=64)
```

3. predict()로 test set 의 anomaly score 를 구하고 iso_result 에 저장해주세요.

```
anomaly_score<-predict(model,newdata=test)
iso_result<-anomaly_score
head(iso_result,10)
```

```
##         3         4         5        13        15        20        23        24
## 0.5797373 0.5831327 0.6141245 0.5861440 0.5627580 0.6031189 0.5809557 0.5793885
##        25        28
## 0.5879789 0.5898567
```

4. 방금 생성한 iso_result 를 0.6 이상이면 1,이하면 0 으로 재범주화 시켜주세요.

```
iso_result<-ifelse(iso_result>=0.6,1,0)
```

문제 9. t-sne 시각화 (https://lovit.github.io/nlp/representation/2018/09/28/tsne/ (https://lovit.github.io/nlp/representation/2018/09/28/tsne/))

1. t-sne 에 대해 간단히 알아보고 perplexity 가 무엇을 의미하는지도 알아보세요!(선택)

높은 차원의 복잡한 데이터를 2차원에 차원 축소하는 방법이다. 낮은 차원 공간의 시각화에 주로 사용하며 차원 축소할 때는 비슷한 구조끼리 데이터를 정리한 상태이므로 데이터 구조를 이해하는 데 도움을 준다.

매니폴드 학습의 하나로 복잡한 데이터의 시각화가 목적이다. 높은 차원의 데이터를 2차원 또는 3차원으로 축소시켜 시각화한다.

t-SNE를 사용하면 높은 차원 공간에서 비슷한 데이터 구조는 낮은 차원 공간에서 가깝게 대응하며, 비슷하지 않은 데이터 구조는 멀리 떨어져 대응된다.

이웃 데이터 포인트에 대한 정보를 보전하려고 한다.

perplexity : 학습에 영향을 주는 점들의 개수를 조절

2. Rtsne 패키지를 호출하세요.
3. Rtsne 함수를 활용해 test set 을 2 차원으로 변형시켜줍니다. (perplexity = 50)
4. 2 차원으로 변형된 테스트 셋을 result 라는 이름의 데이터프레임으로 만들어주세요.
5. result 에 문제 8-4 에서 만든 iso_result 를 새로운 열로 추가해주세요.

```
library(Rtsne)
```

```
## Warning: 패키지 'Rtsne'는 R 버전 4.1.3에서 작성되었습니다
```

```
test_result<-Rtsne(test,perplexity = 50,check_duplicates = FALSE)
result<-data.frame(v1=test_result$Y[,1],v2=test_result$Y[,2]) %>% mutate(iso_result)
```

문제 10. k-means 를 활용한 이상치 탐지

1. “factoextra”, “cluster”, “gridExtra” 패키지를 호출하세요.

```
library(factoextra)
```

```
## Warning: 패키지 'factoextra'는 R 버전 4.1.3에서 작성되었습니다
```

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

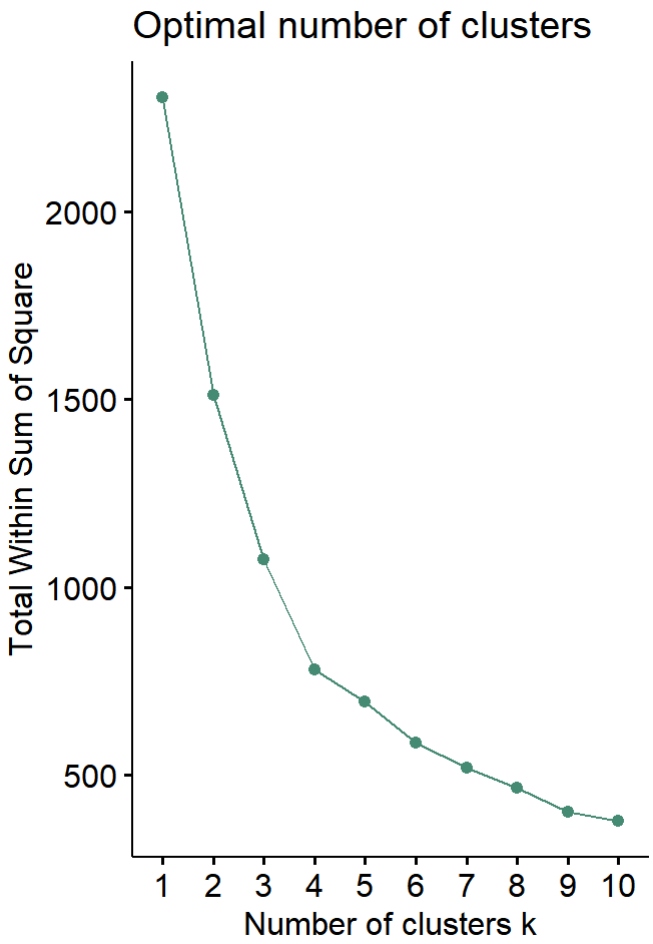
```
library(cluster)
library(gridExtra)
```

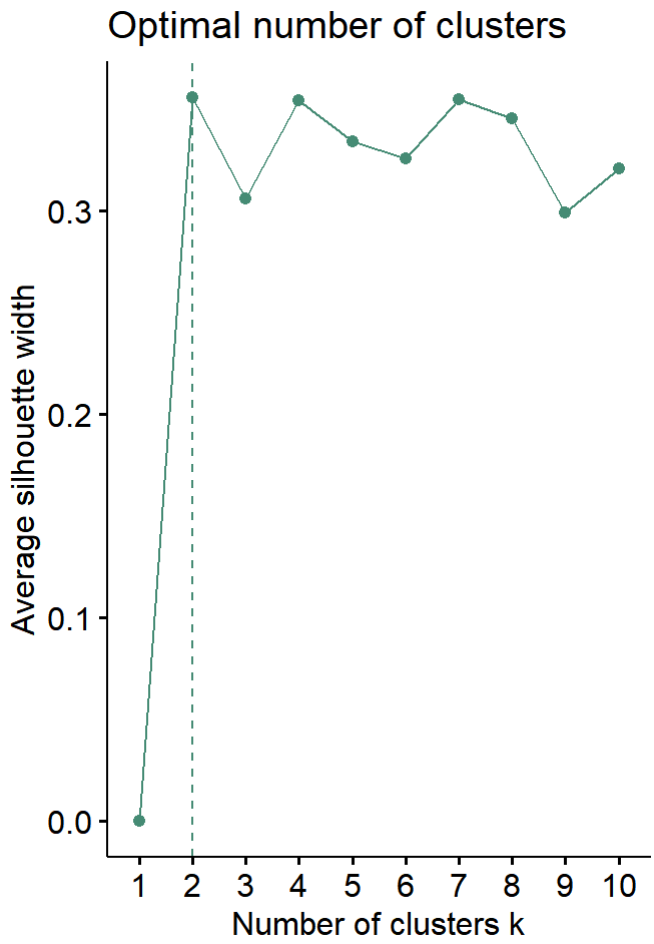
```
##
## 다음의 패키지를 부착합니다: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

2. 아래의 플랏을 그려보고 k 의 개수를 몇으로 해야할지 말해보세요. (색상은 aquamarine4, fviz_nbclust 활용)

```
wwsplot<-fviz_nbclust(train,linecolor='aquamarine4',method='wss',kmeans)
silplot<-fviz_nbclust(train,linecolor='aquamarine4',method='silhouette',kmeans)
grid.arrange(wwsplot,silplot, ncol=2)
```





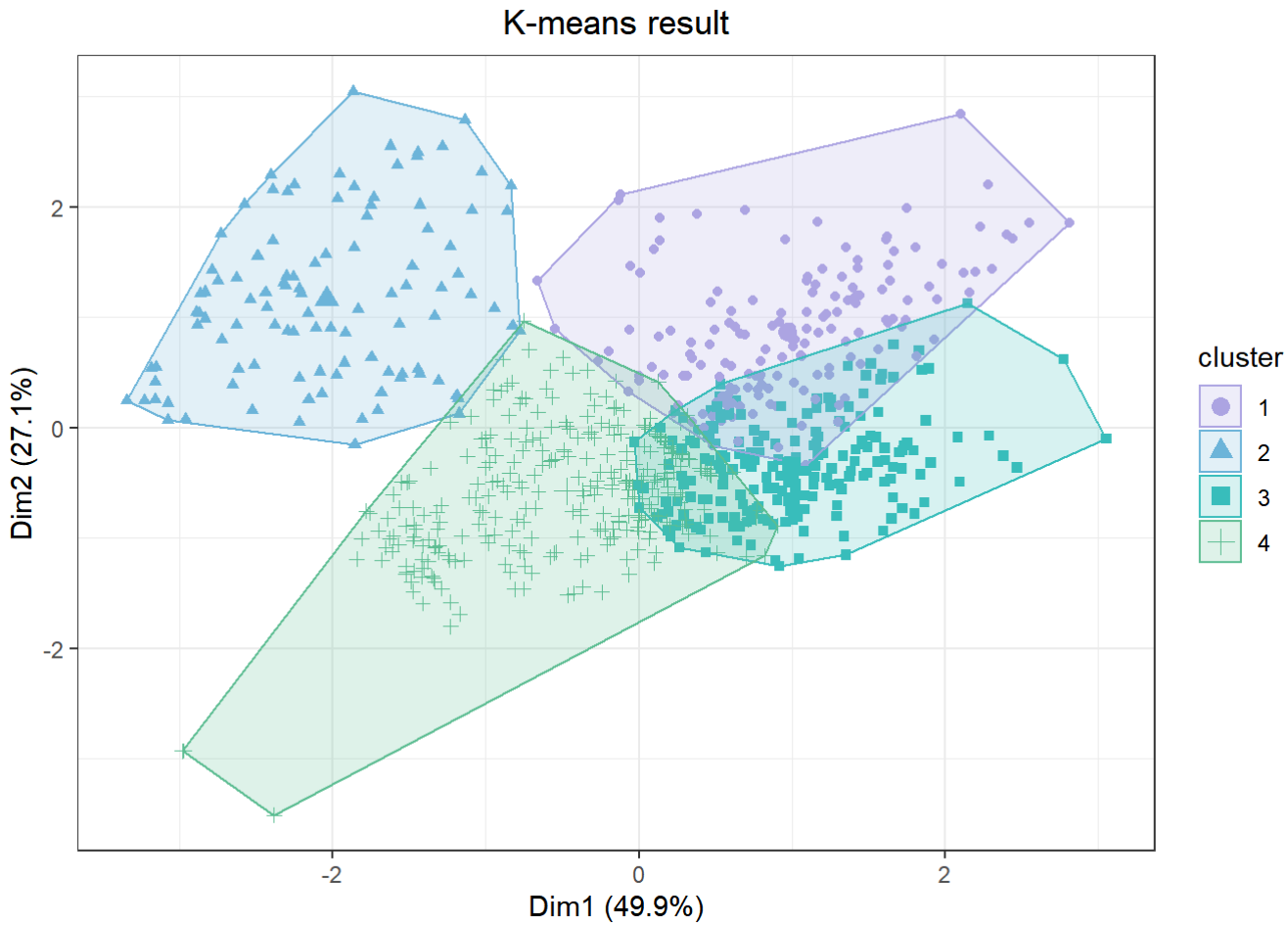
3. k=4, iter.max=50,nstart=1 로 설정한 뒤 train 을 활용해 클러스터링을 진행한 후 아래와 같이 결과를 표현해보세요.(seed 는 2829)

```
set.seed(2829)

h4 <- hcl.colors(4, palette = "cold")

kmcluster<-kmeans(train,4,iter.max=50,nstart=1)

fviz_cluster(kmcluster,train,stand=T,geom='point')+
  scale_fill_manual(values=h4)+
  scale_color_manual(values=h4)+
  theme_bw()+
  ggtitle("K-means result")+
  theme(plot.title=element_text(hjust=0.5))
```



4. 위에서 train 으로 구한 cluster 를 바탕으로 test 에 cluster 를 배정합니다.

```
library("fdm2id")
```

```
## Warning: 패키지 'fdm2id'는 R 버전 4.1.3에서 작성되었습니다
```

```
## 필요한 패키지를 로딩중입니다: arules
```

```
## Warning: 패키지 'arules'는 R 버전 4.1.3에서 작성되었습니다
```

```
## 필요한 패키지를 로딩중입니다: Matrix
```

```
##
## 다음의 패키지를 부착합니다: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## 다음의 패키지를 부착합니다: 'arules'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
## 필요한 패키지를 로딩중입니다: mclust
```

```
## Warning: 패키지 'mclust'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Package 'mclust' version 5.4.9
## Type 'citation("mclust")' for citing this R package in publications.
```

```
##
## 다음의 패키지를 부착합니다: 'mclust'
```

```
## The following object is masked from 'package:VIM':
##
##     diabetes
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
## 필요한 패키지를 로딩중입니다: nnet
```

```
## 필요한 패키지를 로딩중입니다: pls
```

```
## Warning: 패키지 'pls'는 R 버전 4.1.3에서 작성되었습니다
```

```
##
## 다음의 패키지를 부착합니다: 'pls'
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
##
## 다음의 패키지를 부착합니다: 'fdm2id'
```

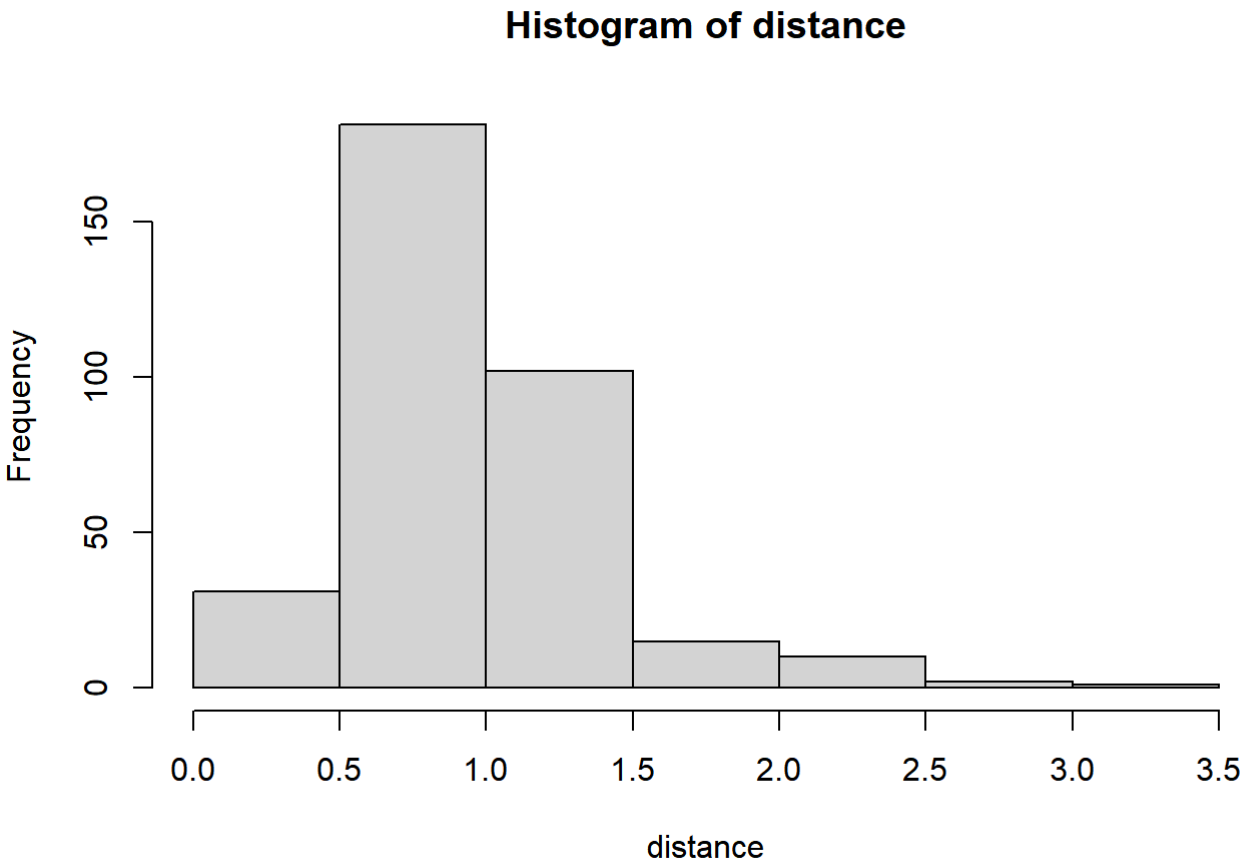
```
## The following objects are masked from 'package:VIM':
##
##     evaluation, wine
```

```
km_result<-predict(kmcluster,data=train,newdata=test)
clust_test<-cbind(test,km_result)
```

5. $centers 를 통해 각 클러스터의 중심점을 출력할 수 있습니다. test 의 행들이 각각 자기가 속한 클러스터의 중심점과의 거리를 구해주세요.

```
center<-kmcluster$centers %>% data.frame
distance<-c()
for (i in 1:nrow(clust_test)){
  testrow<-clust_test[i,c(1:3)]
  matr<-rbind(testrow,center[clust_test$km_result[i],])
  long<-dist(matr,method='euclidian')
  distance<-c(distance,long)
}

hist(distance)
```



6. 중심점과의 거리가 3Q+1.5IQR 이 넘는 점을 outlier 로 지정해주세요
7. km_result 를 test 에서 outlier 로 지정되었으면 1, 아니면 0 인 이진변수로 변경해주세요.

```
outlier_val<-quantile(distance,0.75)+1.5*IQR(distance)
outlier_index<-which(distance>outlier_val)
for (i in 1:length(km_result)){
  if(i %in% outlier_index==T){km_result[i]<-1}else{
    km_result[i]<-0
  }
}
```

문제 11. DBSCAN 으로 이상치 탐지하기

1. fpc 패키지 호출

```
library(fpc)
```

```
## Warning: 패키지 'fpc'는 R 버전 4.1.3에서 작성되었습니다
```

```
##
## 다음의 패키지를 부착합니다: 'fpc'
```

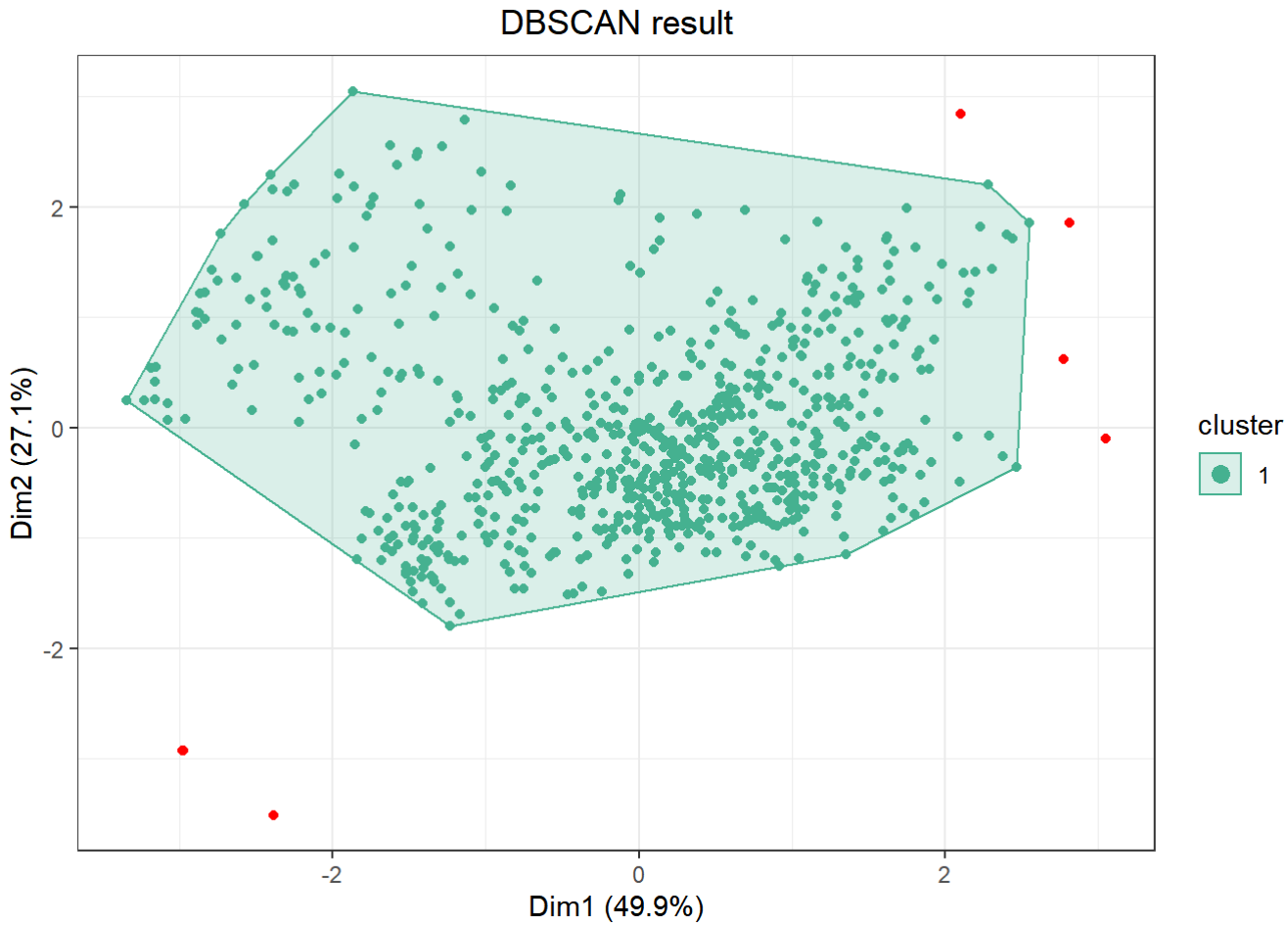
```
## The following object is masked from 'package:fdm2id':
##
##     confusion
```

2. train 으로 eps = 0.7 ,MinPts = log(1139)으로 클러스터링 해주세요.(dbscan())

```
set.seed(2829)
db<-dbscan(train,eps = 0.7 ,MinPts = log(1139))
```

3. 결과를 시각화 하는데 이때, 이상치(클러스터 0)으로 분류된 값은 빨간색으로 표시해주세요.

```
col <-ifelse(db$cluster==1,'#45b190','red')
fviz_cluster(db,train,stand=T,geom='point',
             outlier.color = 'red',
             palette=col)+
  theme_bw()+
  ggtitle("DBSCAN result")+
  theme(plot.title=element_text(hjust=0.5))
```



4. predict.dbscan()으로 test 셋에도 위에서 만든 클러스터에 배정되도록 하고 outlier 인 경우 1,아닌 경우 0 으로 재범주화 한 뒤 db_result 로 저장해주세요.

```
db_predict<-predict.dbscan(db,data=train,newdata = test)
db_result<-ifelse(db_predict==0,1,0)
```

문제 12. tsne 시각화

1. result 에 km_result 와 db_result 를 새로운 열로 추가해주세요

```
result %<>% mutate(km_result,db_result) %>%
  mutate_at(vars(iso_result,km_result,db_result),as.factor)
```

2. 아래와 같이 이상치는 “#B5615B” 정상값은”#E8E5E5”로 결과를 시각화해보세요.

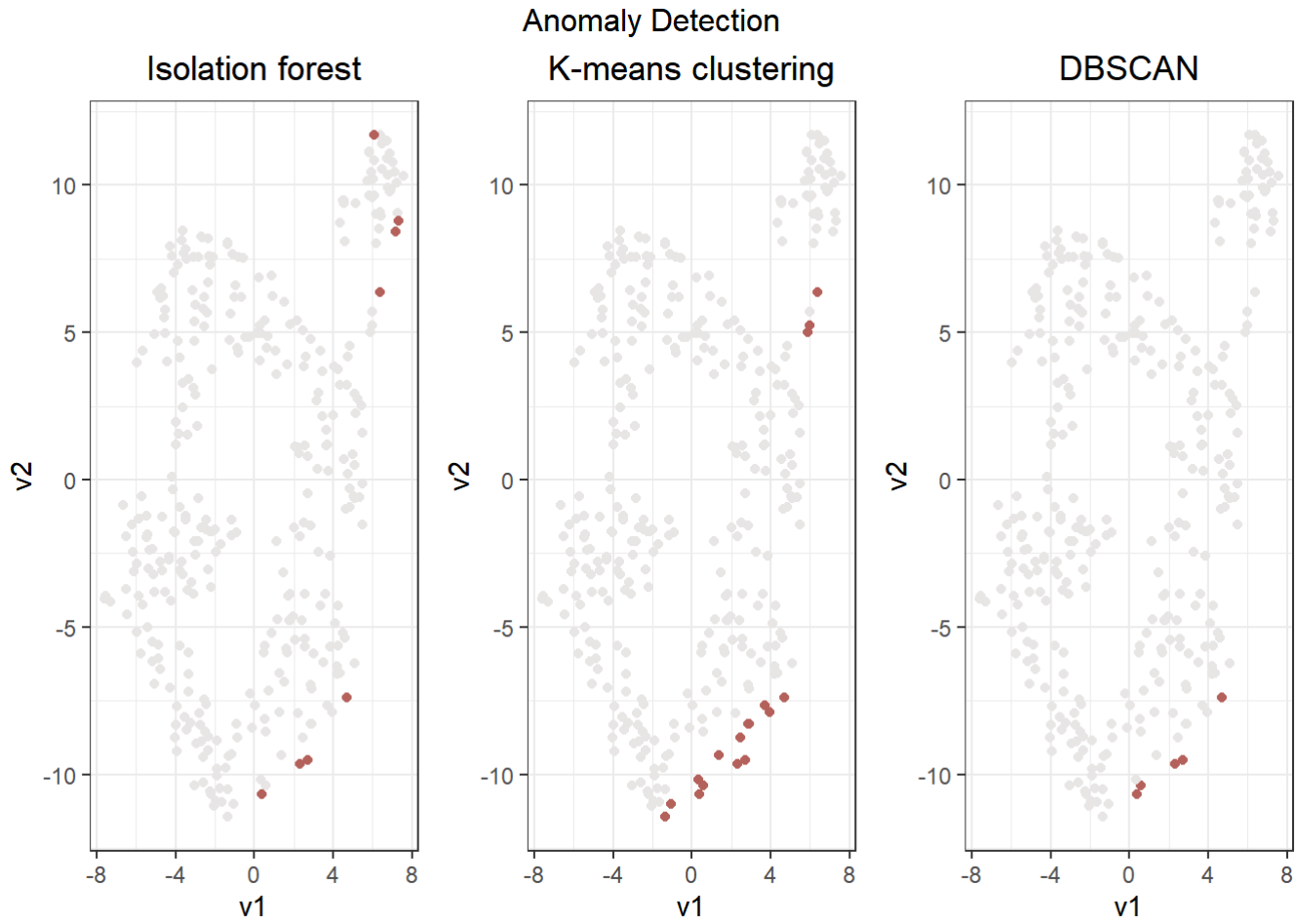
```
col<-c('1'='#B5615B','0'='#E8E5E5')


iso<-ggplot(result,(aes(v1,v2,color=iso_result)))+
  geom_point()+
  ggtitle("Isolation forest")+
  scale_color_manual(values=col)+
  theme_bw()+
  theme(plot.title=element_text(hjust=0.5),
        legend.position="none")


kms<-ggplot(result,(aes(v1,v2,color=km_result)))+
  geom_point()+
  ggtitle("K-means clustering")+
  scale_color_manual(values=col)+
  theme_bw()+
  theme(plot.title=element_text(hjust=0.5),
        legend.position="none")


dbsc<-ggplot(result,(aes(v1,v2,color=db_result)))+
  geom_point()+
  ggtitle("DBSCAN")+
  scale_color_manual(values=col)+
  theme_bw()+
  theme(plot.title=element_text(hjust=0.5),
        legend.position="none")

grid.arrange(iso,kms,dbsc,ncol=3,top='Anomaly Detection')
```



part2. 지도시각화

```
library(ggmap)
```

```
## Warning: 패키지 'ggmap'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

```
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
##
## 다음의 패키지를 부착합니다: 'ggmap'
```

```
## The following object is masked from 'package:magrittr':
##
##     inset
```

```
# data <- fread("part2_data.csv", encoding = "UTF-8")
#cood<-geocode(data$도로명주소,source='google')
#data<-cbind(data,cood)
#data<-na.omit(data) na가 나온 행은 제거
#write.csv(data, file = "data_api.csv", row.names = FALSE)

data_api<-fread( "data_api.csv")
```

문제 3. 시/도를 의미하는 addr1 변수와 시/군/구를 의미하는 addr2 변수를 만들어주세요.(str_sub, str_extract)

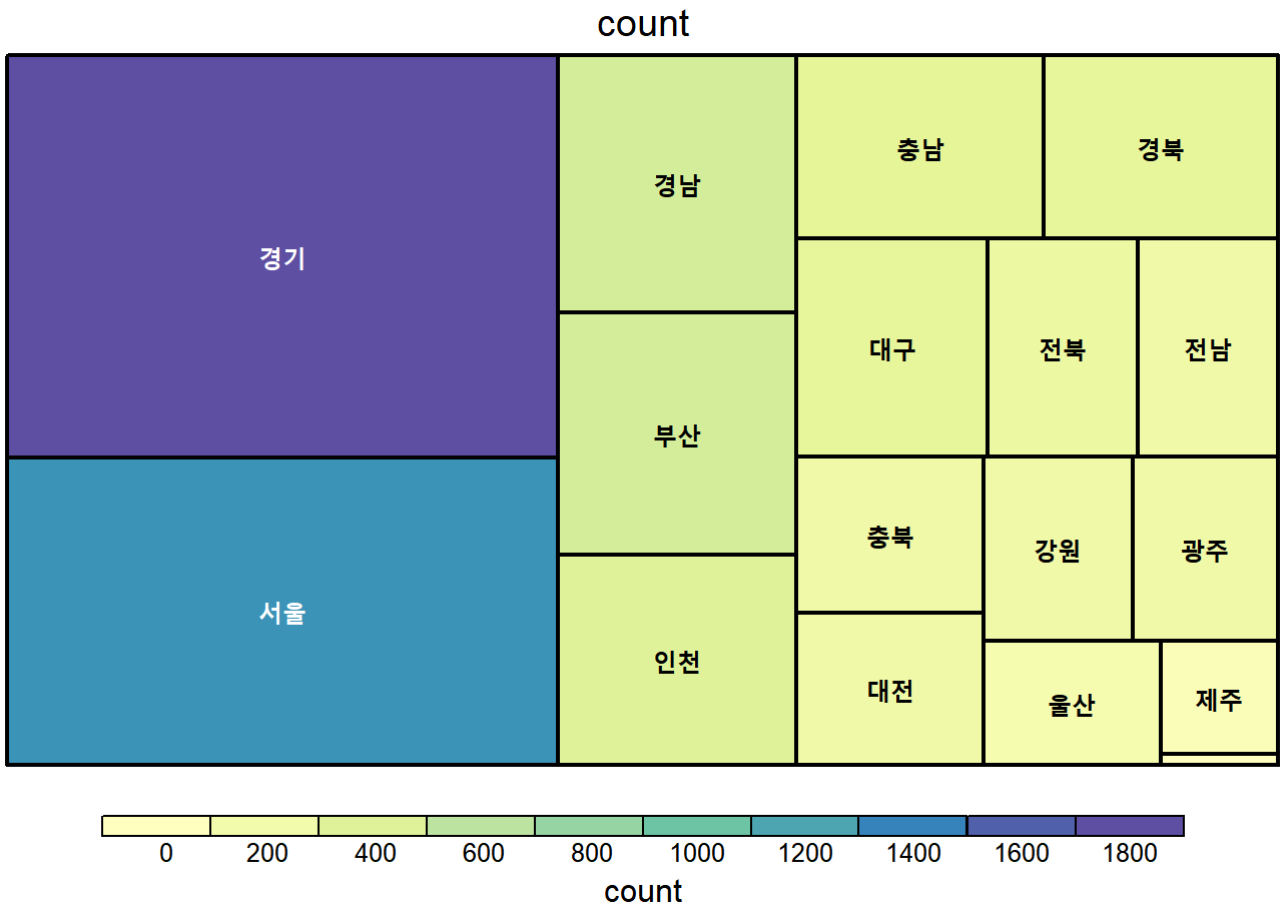
```
x<-str_split_fixed(data_api$도로명주소,' ',3)
data_api %<>% mutate(addr1=x[,1],addr2=x[,2])
data_api$addr2[which(data_api$addr1=='세종')]='세종'
#data$addr2<-str_extract(data$도로명주소,"WWw+[시|군|구]")
```

문제 4. tree map 시각화. (count 는 addr1 에 따른 복권상점 개수)

```
library(treemap)
```

```
## Warning: 패키지 'treemap'는 R 버전 4.1.3에서 작성되었습니다
```

```
count<- data_api %>% group_by(addr1) %>% summarise(count=n())
treemap<-treemap(count,index=c("addr1"),
                 vColor="count",
                 type='value',
                 vSize='count',
                 palette='Spectral')
```



문제 5. 지도시각화

1.

```
library(sf)
```

```
## Linking to GEOS 3.9.1, GDAL 3.2.1, PROJ 7.2.1; sf_use_s2() is TRUE
```

```
library(rgdal)
```

```
## Warning: 패키지 'rgdal'는 R 버전 4.1.3에서 작성되었습니다
```

```
## 필요한 패키지를 로딩중입니다: sp
```

```
## Warning: 패키지 'sp'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Please note that rgdal will be retired by the end of 2023,
## plan transition to sf/stars/terra functions using GDAL and PROJ
## at your earliest convenience.
##
## rgdal: version: 1.5-28, (SVN revision 1158)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 3.2.1, released 2020/12/29
## Path to GDAL shared files: C:/Users/shn20/Documents/R/win-library/4.1/rgdal/gdal
## GDAL binary built with GEOS: TRUE
## Loaded PROJ runtime: Rel. 7.2.1, January 1st, 2021, [PJ_VERSION: 721]
## Path to PROJ shared files: C:/Users/shn20/Documents/R/win-library/4.1/rgdal/proj
## PROJ CDN enabled: FALSE
## Linking to sp version:1.4-6
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,
## use options("rgdal_show_exportToProj4_warnings"="none") before loading sp or rgdal.
## Overwritten PROJ_LIB was C:/Users/shn20/Documents/R/win-library/4.1/rgdal/proj
```

```
map_korea<-readOGR('TL_SCCO_CTPRVN.shp')
```

```
## Warning in OGRSpatialRef(dsn, layer, morphFromESRI = morphFromESRI, dumpSRS =
## dumpSRS, : Discarded datum International_Terrestrial_Reference_Frame_2000 in
## Proj4 definition: +proj=tmerc +lat_0=38 +lon_0=127.5 +k=0.9996 +x_0=1000000
## +y_0=2000000 +ellps=GRS80 +units=m +no_defs
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "C:WUsersWshn20WDesktopW작업물W통계분석학회W2022-1학기W패키지 과제W[22학년도 1학기]
클린업 3주차 패키지WTL_SCCO_CTPRVN.shp", layer: "TL_SCCO_CTPRVN"
## with 17 features
## It has 3 fields
```

```
map_korea<-map_korea %>% spTransform(CRS('+proj=longlat'))
map_korea<-fortify(map_korea)
```

```
## Regions defined for each Polygons
```

2. map_korea 를 활용해 아래의 지도를 그려보세요.

```
library(RColorBrewer)
col<-brewer.pal(12,"Paired")
col<-colorRampPalette(col)(20)
ggplot()+
  geom_polygon(data=map_korea,aes(long,y=lat,group=group),fill='white',
               color='grey',size=0.1)+
  geom_point(data=data_api,aes(lon,lat,color=addr1),size=0.5,alpha=0.5)+
  scale_color_manual(values=col,
                     name='')+
  ggtitle("전국 복권판매점 분포")+
  theme_classic()+
  theme(plot.title=element_text(face='bold',hjust=0.5,size=15),
        legend.text=element_text(size=7))
```