# package1

jeonhyolim

2021 3 11

## 1 주차 패키지_ 27 기 전효림

getwd() setwd("C:/Users/전효림/Documents") getwd() #chapter1 #문제 0
install.packages('plyr') library(plyr) install.packages('dplyr') library(dplyr)
install.packages('tidyverse') library(tidyverse) install.packages('data.table')
library(data.table)

setwd("C:/Users/전효림/Desktop/5 학기/피셋/패키지문제/1 주차") getwd() df <-
fread("data.csv", stringsAsFactors = FALSE, data.table = FALSE)

#문제 1 str(df)

length(df$sex$)unique(df$sex) length(df$age$)unique(df$age)
length(df$country$)unique(df$country) length(df$province$)unique(df$province)
length(df$city$)unique(df$city) length(df$confirmed_date$)unique(df$confirmed_date)
length(df$state$)unique(df$state)

## 문제 2-1.

df <- na.omit(df) str(df)

## 문제 2-2

df[df==""]<-NA df<-df %>% filter(!is.na(sex)) df<-df %>% filter(!is.na(age)) df<-df %>%
filter(!is.na(city)) str(df)

length(df$sex$)unique(df$sex) length(df$age$)unique(df$age)
length(df$country$)unique(df$country) length(df$province$)unique(df$province)
length(df$city$)unique(df$city) length(df$confirmed_date$)unique(df$confirmed_date)
length(df$state$)unique(df$state)

## 문제 3

df_1 <- subset(df,df$country=="Korea") str(df_1) unique(df_1$country) #확인용

df_2 <-df_1[,-c(4)] str(df_2)

## 문제 4

df_2$province = plyr::revalue(df_2$province,c("서울"="서울특별시","부산"="부산광역시", "대구" = "대구광역시", "인천" = "인천광역시", "대전" = "대전광역시" , "세종" = "세종특별자치시", "울산"= "울산광역시", "제주도"= "제주특별자치도")) str(df_2)

## 문제 5

df_2$confirmed_date = as.Date(df_2$confirmed_date) str(df_2)
#as.Date(df_2$confirmed_date) #str(df_2)

## 문제 6

df_2= df_2 %>% group_by(confirmed_date) %>% mutate(confirmed_number = n()) %>% ungroup() %>% as.data.frame() data %>% head #summary(df_2) #confirmed_number <- table(df_2$confirmed_date) #str(df_2)

## 문제 7

df_2$wday = NAdays = weekdays(df_2$confirmed_date) for (j in 1:nrow(df_2)){ if (days[j] %in% c("토요일","일요일")){ df_2$wday[j]="주말" } else {df_2$wday[j]="주중"} } data %>% head #my_date_list <- as.Date(df_2$confirmed_date) #weekdays(my_date_list) #day_levels <- c("일요일", "월요일", "화요일", "수요일", "목요일", "금요일", "토요일") #factor(weekdays((my_date_list)), levels=day_levels, ordered=TRUE) #as.numeric(factor(weekdays((my_date_list)), levels=day_levels, ordered=TRUE))

#wday <- ifelse(weekdays(as.Date(df_2$confirmed_date))== "토요일", "주말", "주중") #str(df_2)

## 문제 8

tapply(df_2$confirmed_date$, df_2$age, summary) data2=df_2 %>% group_by(age,confirmed_date) %>% summarise(nums=n()) %>% as.data.frame() tapply(data2$nums$, data2$age, summary)

## chapter2. 시각화

## 문제 1-1

install.packages('ggplot2') library(ggplot2)

class(df_2) str(df_2) df_2 %>% ggplot(aes(x=confirmed_date,y=confirmed_number))+ geom_line(color="lightblue")+ ggtitle("코로나 확진자수 추이-국내인 기준")+ geom_point(aes(x =df_2[df_2$confirmed_number$ == max(df_2$confirmed_number),6] %>% unique(), y = max(df_2$confirmed_number) ), color="navy") + annotate("text", x=df_2[df_2$confirmed_number==max(df_2$confirmed_number),6]confirmed_number), label=paste0(df_2[df_2$confirmed_number$ == max(df_2$confirmed_number),6] %>% unique(),"(",max(df_2$confirmed_number),"명)") ,color = "navy",hjust = 1)+ theme_classic()+ theme(plot.title = element_text(hjust = 0.5,face="bold"))

#문제 1-2 df_2 %>% ggplot(aes(x=confirmed_date, y=confirmed_number))+geom_line(stat='identity')+facet_wrap(~province,ncol=4)

## 문제 2

df_2 %>% ggplot(aes(x=confirmed_number, y=reorder(province,confirmed_number), fill=state)) +geom_bar(stat='identity')+geom_col()

+facet_wrap(~state)

class(df_2) str(df_2)

## 문제 3

library(lattice)

## df_2 %>% ggplot(aes(x=age, y=confirmed_number/length(unique(confirmed_date)), fill=age)) +geom_boxplot()

#일별 확진자 수= 전체 확진자수 / 전체 기간 length(unique(df_2$confirmed_date))

data2 %>% group_by(age) %>% ggplot(aes(x=age, y=nums,color=age,fill=age))+ geom_boxplot(alpha=0.6,outlier.shape = NA)+ ylab("월단위 확진자수")+ theme_classic()

## 문제 3-2

aov(nums ~ ., data = data2)

## Chapter3

## 문제 1

install.packages('MASS') library(MASS) install.packages('corrplot') library(corrplot) install.packages('caret') library(caret) install.packages('MLmetrics') library(MLmetrics)

data(Boston) dim(Boston)

head(Boston)

glimpse(Boston)

summary(Boston)

sum(is.na(Boston))

sum(duplicated(Boston))

corrplot(cor(Boston), method = "number", type = "upper", diag = FALSE)

#문제 2 Boston %>% gather(key, val, -medv) %>% ggplot(aes(x = val, y = medv)) + geom_point() + stat_smooth(method = "lm", se = TRUE, col = "blue") + facet_wrap(~key, scales = "free") + theme_gray() + ggtitle("Scatter plot of dependent variables vs Median Value (medv)")

table(Boston$chas)

#문제 3 set.seed(1234) train_index = createDataPartition(Boston$medv, p = 0.7)$train = $Boston[train_index$Resample1,] test = Boston[-train_index$Resample1,]

#문제 3-2

model=lm(medv~.,data=train) pred=predict(model, test[,-14]) RMSE(pred, test$medv)

## 문제 3-3

## RMSE(평균제곱근오차)는정량적 데이터를 예측할 때 모델의 오차를 측정하는 표준 방법이다.RMSE 가 예측 된 값의 벡터와 관찰 된 값의 벡터 사이의 일종의 (정규화 된) 거리로 생각 될 수 있어서 의미가 있다.

## 문제 4

summary(model) cof = model$coefficients$cols= rownames(cof) colnames(cof)= c("values","val")

cof %>% ggplot(aes(x=values, y=reorder(val,values),fill=values,color=values))+ geom_bar(stat = 'identity',alpha=0.3)+ theme_classic()+ xlab("value")+ ylab("intercept and variable")+ scale_color_gradient2(low="blue", mid="yellow",high="red")+ scale_fill_gradient2(low="blue", mid="yellow",high="red")+ theme(legend.position = 'none') + geom_text(aes(label = round(values,2)),color="black",position=position_stack(0.5))