library(pdftools)

library(tesseract)

df=data.frame(matrix('',ncol = 0,nrow = 20))

for (i in 1:48) {

pth=paste('C:/Users/Shabrez/Desktop/E10/a (',').pdf',sep=as.character(i))

print(pth)

pdf\_convert(

pth,

format = "png",

pages = 2,

filenames = NULL,

dpi = 72,

antialias = TRUE,

opw = "",

upw = "",

verbose = TRUE

)

}

library(magrittr)

library(tidyverse)

library(magick)

for (i in 1:48) {

pth=paste('C:/Users/Shabrez/Documents/a (',')\_2.png',sep=as.character(i))

raw\_img=image\_read(pth)

image\_ggplot(raw\_img)

raw\_img %>%

image\_crop(geometry\_area(350, 0,350, 10)) %>%

ocr()

raw\_img %>%

image\_quantize(colorspace = "gray") %>%

image\_ggplot()

fuzz\_fun <- function(fuzz){

raw\_img %>%

image\_quantize(colorspace = "gray") %>%

image\_transparent(color = "white", fuzz=fuzz) %>%

image\_background("white") %>%

image\_crop(geometry\_area(350, 0,350, 10))

}

fuzz\_fun(20)

combo\_fuzz <- c(

fuzz\_fun(100),

fuzz\_fun(100),

fuzz\_fun(100),

fuzz\_fun(100)

) %>%

image\_append(stack = TRUE)

image\_ggplot(combo\_fuzz)

no\_grid <- raw\_img %>%

image\_quantize(colorspace = "gray") %>%

image\_transparent(color = "white", fuzz=20) %>%

image\_background("white")

image\_ggplot(no\_grid)

no\_grid %>%

image\_negate() %>%

image\_ggplot()

no\_grid %>%

image\_negate() %>% # negate

image\_morphology(method = "Thinning", kernel = "Rectangle:20x1") %>%

image\_negate() %>% # back to white

image\_ggplot()

no\_grid %>%

image\_crop(geometry\_area(350, 0,350, 10)) %>%

image\_ggplot()

no\_grid\_crop <- no\_grid %>%

image\_crop(geometry\_area(350, 0,350, 10))

no\_grid\_crop %>%

image\_ggplot()

no\_grid\_crop %>%

image\_ocr()

num\_only <- tesseract::tesseract(

options = list(tessedit\_char\_whitelist = c(".0123456789 "))

)

no\_grid %>%

image\_quantize(colorspace = 'gray') %>%

image\_threshold() %>%

image\_crop(geometry\_area(350, 0,350, 10)) %>%

ocr(engine = num\_only)

combo <- tesseract::tesseract(

options = list(

tessedit\_char\_whitelist = paste0(

c(letters, LETTERS, " ", ".0123456789 (-)"), collapse = "")

)

)

raw\_text <- no\_grid %>%

image\_quantize(colorspace = "gray") %>%

image\_transparent("white", fuzz = 22) %>%

image\_background("white") %>%

image\_threshold() %>%

image\_crop(geometry\_area(300, 0,400, 10)) %>%

ocr(engine = combo)

raw\_tibble <- raw\_text %>%

str\_split(pattern = "\n") %>%

unlist() %>%

tibble(data = .)

raw\_tibble$data=gsub(')','',as.character(raw\_tibble$data))

raw\_tibble$data=gsub('-','',as.character(raw\_tibble$data))

col1=raw\_tibble[1:7,]

raw\_tibble=raw\_tibble[-(1:9),]

raw\_tibble=raw\_tibble[-(11:13),]

col1=col1[-(1:2),]

col1=sub('.\*? ','',col1$data)

col1=data.frame(col1)

col1$col1

check=separate(raw\_tibble,data ,into = c('retail1','retail2','direct'),sep = '\\ ',extra = 'merge')

library(tidyverse)

col1=col1 %>% add\_row(col1=check$retail2)

df=data.frame(df,col1)

}