18f0371 talha Zafar

```
raw_img %>%
  image_crop(geometry_area(350, 0,350, 10)) %>%
  ocr()
```

```
## [1] "ye May 5, 2019\n|\n[I =\nRe/Liter\n[| 73.407\nncease 72.001\n720\n73a0
| 66.06 |\n7326 |\nMOGAS E510 Gasoline\nRetail_| Retail | Direct\n73.26 |
73.26\n[329 | ose| 056\n(76.69 | 73.80 | 73.80\n| SeR=oee | EOE |=\n|_76.69 |
73.80 | 73.80\n_ Ser ber ee| 264\ni007] a6 | 809\nite ae |\n[ez | 15.39)
1539\n108.42 Lasse 1085 |\n@sn\nper by age. [231%\n"
```

Including Plots

```
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))
}
```
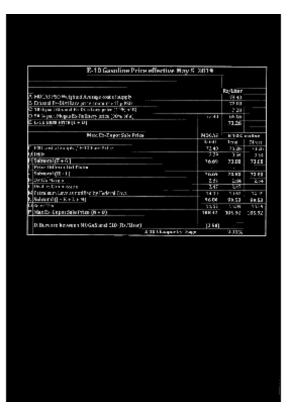
```
fuzz_fun(20)
```

	MOGAS	E-10 Gasoline	
		Rs/Liter	
		73.40	
		72.00	
		7.20	
	73.40	66.06	
		73.26	
	MOGAS	**E-10 Gasoline**	
	Retail	Retail	Direct
	73.40	73.26	73.26
	3.29	0.54	0.54
	76.69	**73.80**	**73.80**
	-	-	-
	76.69	**73.80**	**73.80**
	2.64	2.64	2.64
	3.47	3.47	
	14.00	10.62	14.09
	96.80	**90.53**	**90.53**
	11.62	15.39	15.39
	108.42	**105.92**	**105.92**
	(2.50)	-	
per by %age		2.31%	

```
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()
```

```r
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()
```

```r
no_grid_crop <- no_grid %>%
  image_crop(geometry_area(350, 0,350, 10))

no_grid_crop %>%
  image_ggplot()
```



```r
no_grid_crop %>%
  image_ocr()
```

```
## [1] "ve May 5, 2019\n—\n[I\nRe [tier\n[7340 |\n[200\n720\nTaD 66.06\n73.26
|\nMOGAS \"~F:40 Castine\nRetail | Reta | Direc\n73.26 | 73.26\n[3297 054
054\n[7669 | 73.20 | 7380\nTt\n[7669 | 73.90 | 73.90\n— Ser der ee|\n—
349\n~~ rae |e\n360 [9053 | 9053\nsae | weet 15.39\n108.42 | 105.92 | 105.92
|\nesa\nper by age 23%\n"
```

```r
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)
```

```
## [1] "5 2019\n73.40\n72.00\n7 320\n730 66.06\n73.26\n0 10\n73407326
73.26\n329 054 054\n76.69 73.80 73.80\n76697380 73.80\n264 264 264\n37
```

```
347\n14.00 1062 1408\n9680 9053 90.53\n11.62 15391530\n408.42 105.92
10592\n50\n. 2314.\n"

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

## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 1 rows [6].

library(tidyverse)
col1=col1 %>% add_row(col1=check$retail2)
col1

##       col1
## 1     7340
## 2    72.00
## 3       20
## 4    66.06
## 5    73.26
## 6     7326
## 7      054
## 8     7320
```

```
## 9      7380
## 10      264
## 11      347
## 12  10627
## 13   9053
## 14   1539
## 15 105.92
```

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