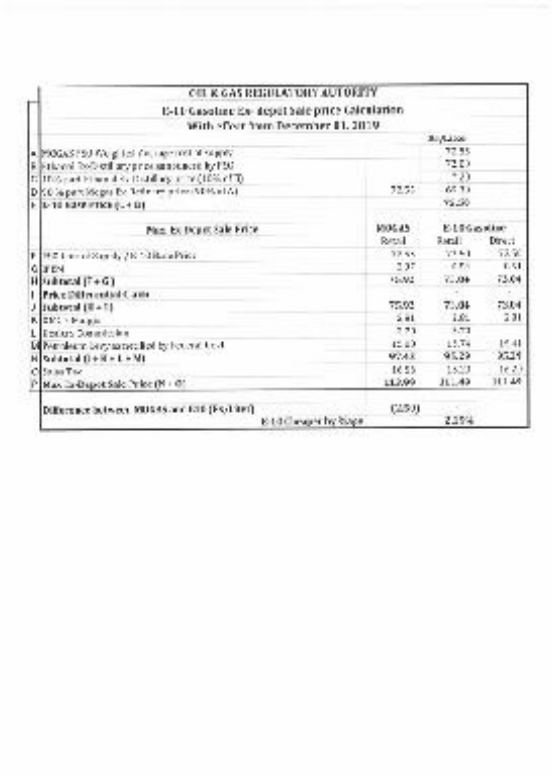
MID 1

17F 8333

11/16/2021

```
#
# for (i in 1:48) {
#   pth=paste('C:/Users/HP/Downloads/E10/pdf (','').pdf',sep=as.character(i))
#   print(pth)
#   pdf_convert(
#     pth,
#     format = "png",
#     pages = 2,
#     filenames = NULL,
#     dpi = 72,
#     antialias = TRUE,
#     opw = "",
#     upw = "",
#     verbose = TRUE
#   )
#
# }

#for (i in 1:48) {
  #pth=paste('pdf (','')_2.png',sep=as.character(i))


raw_img=image_read('pdf (1)_2.png')

image_ggplot(raw_img)
```



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

```
## [1] "UTORITY\nice Calculation\nr
01,2019\nRs/Liter\n7255\n72.00\n7.20\n7255 6530\n72.50\nMOGAS _E-
10Gasoline\nRetall Retail Direct\n755 72507250\n337054 _OS4\n7592 73:04
73.08\n7592 73:04 73.08\n281 2e1 281\n370 370\n15001574 19.48\n9743 95.29
95.29\n1656 1620-1620\n11399 11149 11149\nsy -\nDy Sage 2.19%\n"
```

```
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)
```

AUTORITY

ice Calculation

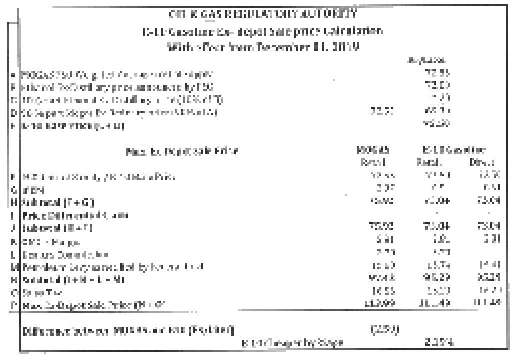
r 01, 2019

		Rs/Liter
		72.55
		72.00
		7.20
	72.55	65.30
		72.50

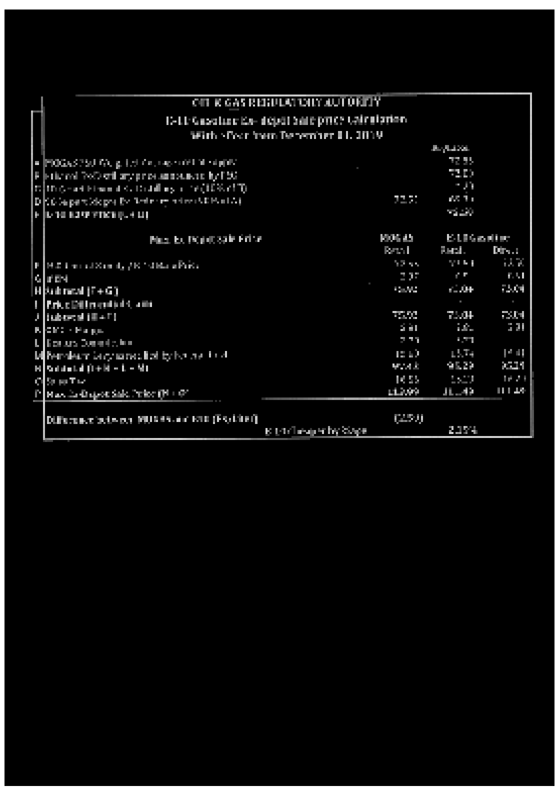
	MOGAS	**E-10 Gasoline**	
	Retail	Retail	Direct
	72.55	72.50	72.50
	3.37	0.54	0.54
	75.92	**73.04**	**73.04**
	-	-	-
	75.92	**73.04**	**73.04**
	2.81	2.81	2.81
	3.70	3.70	
	15.00	15.74	19.44
	97.43	**95.29**	**95.29**
	16.56	16.20	16.20
	113.99	**111.49**	**111.49**
	(2.50)	-	
by %age		**2.19%**	

```
combo_fuzz <- c(
  fuzz_fun(100),
  fuzz_fun(100),
  fuzz_fun(100),
  fuzz_fun(100)
) %>%
  image_append(stack = TRUE)
image_ggplot(combo_fuzz)
```

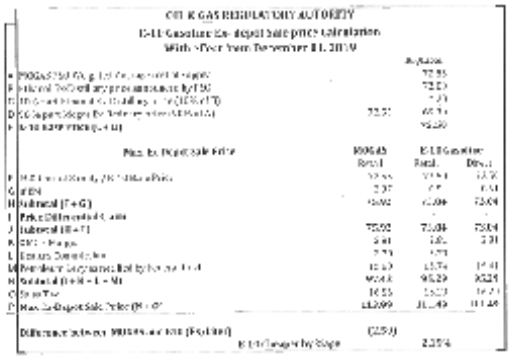
```
no_grid <- raw_img %>%
  image_quantize(colorspace = "gray") %>%
  image_transparent(color = "white", fuzz=20) %>%
  image_background("white")

image_ggplot(no_grid)
```



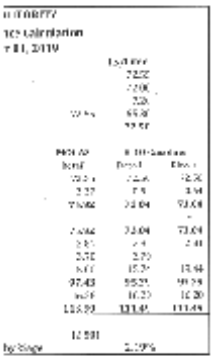
```r
no_grid %>%
  image_negate() %>%
  image_ggplot()
```



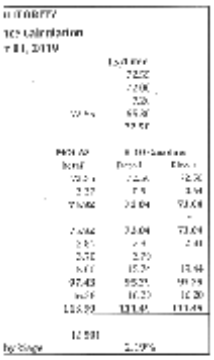
```
no_grid %>%
  image_negate() %>% # negate
  image_morphology(method = "Thinning", kernel = "Rectangle:20x1") %>%
  image_negate() %>% # back to white
  image_ggplot()
```



```
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()
```

```
## [1] "UTORITY\nice Calculation\nr01, 2019\nRs/Liter\n7255\n7200\n720\n7255
6530\n72.80\nMOGAS —_E-10Gasoline\nRetall Retail Direct\n7255 7250-7250\n337
0st OSH\n7592 73:08 73.08\n7592 78.04 73.08\n281 2e1 281\n370
370\n150015741948\n9743 (95.29 95.29\n1656 16.20 1620\n11399 11149
11149\nsy\nDy Mage 2.19%\n"
```

```
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] "01 2019\n72.00\n758 65.30\n7250\n 10\n758 72807250\n337 5 088\n7592
7308 73.04\n7592 73.08 73.08\n281 21 2\n330 370\n00 15741944\n9743 98.29
9529\n1656 16.20 6.20\n11399 1114911149\n250\n2.199\n"
```

```
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 = num_only)

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

## [1] "65.30"       "7250"        "10"          "75872807250" "5 088"

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

## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 6 rows [2,
3, 6,
## 7, 8, 9].

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

##           col1
## 1        65.30
## 2         7250
## 3           10
## 4  75872807250
## 5        5 088
## 6          201
## 7          370
## 8     15781944
## 9        98.29
## 10       16.20
## 11  1114911149
## 12        <NA>
## 13        <NA>
## 14        <NA>

#}
```

R Markdown