RELATÓRIO DE ANÁLISE DESCRITIVA

FORTALEZA
2018

1 Análise das despesas

A seguir temos o TOP 100 das Despesas que mais aparecem no dataset:

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  6621231 353.7   10018085 535.1  6621231 353.7
## Vcells 76979868 587.4  127031987 969.2 76979868 587.4
```

As medidas de resumo mostram que claramente há valores inválidos e outliers nos dados.

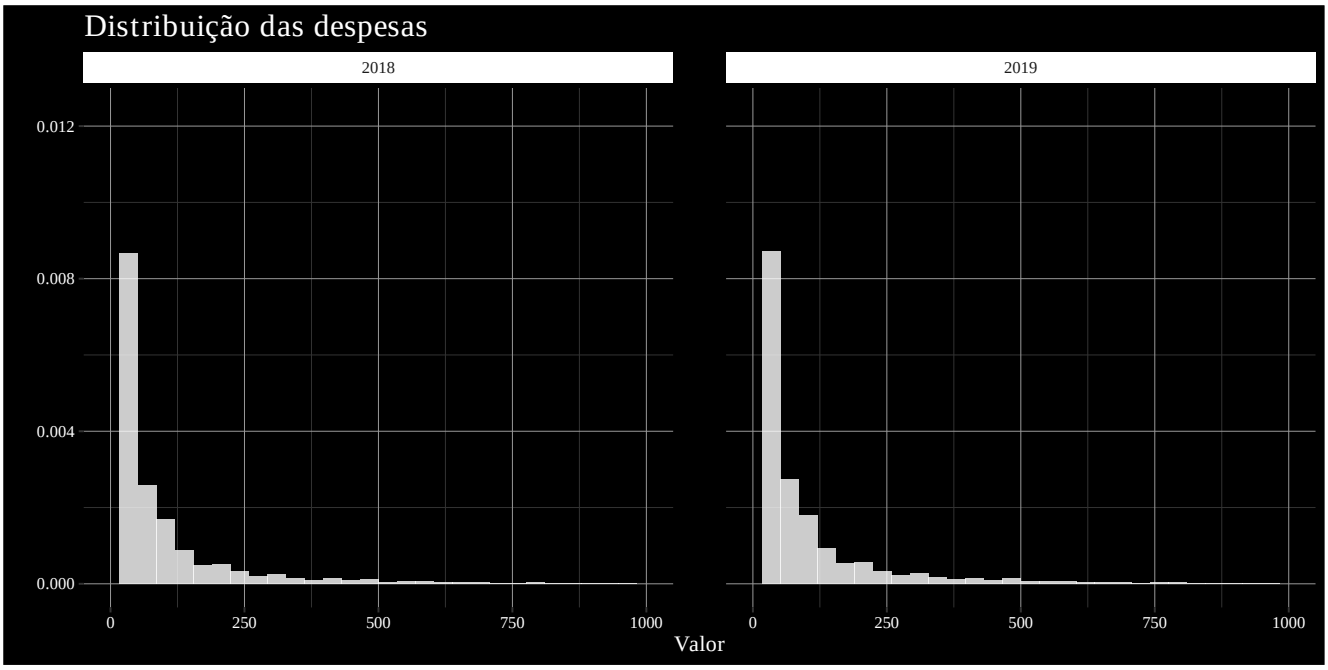
```
## Error in cat(x, file = file, sep = c(rep.int(sep, ncolumns - 1), "\n"), :  objeto 'textoDesp'
não encontrado
```

```
Despesas %>%
  select(ano,Valor)%>%
  split(.$ano) %>%
  map(summary)

## $`2018`
##       ano           Valor
##  Min.   :2018   Min.   : -10000
##  1st Qu.:2018   1st Qu.:      9
##  Median :2018   Median :     22
##  Mean   :2018   Mean   :    133
##  3rd Qu.:2018   3rd Qu.:     60
##  Max.   :2018   Max.   :6000000
##
## $`2019`
##       ano           Valor
##  Min.   :2019   Min.   :          -15591
##  1st Qu.:2019   1st Qu.:              10
##  Median :2019   Median :              25
##  Mean   :2019   Mean   :     12899026578
##  3rd Qu.:2019   3rd Qu.:              69
##  Max.   :2019   Max.   :9000000000000000
```

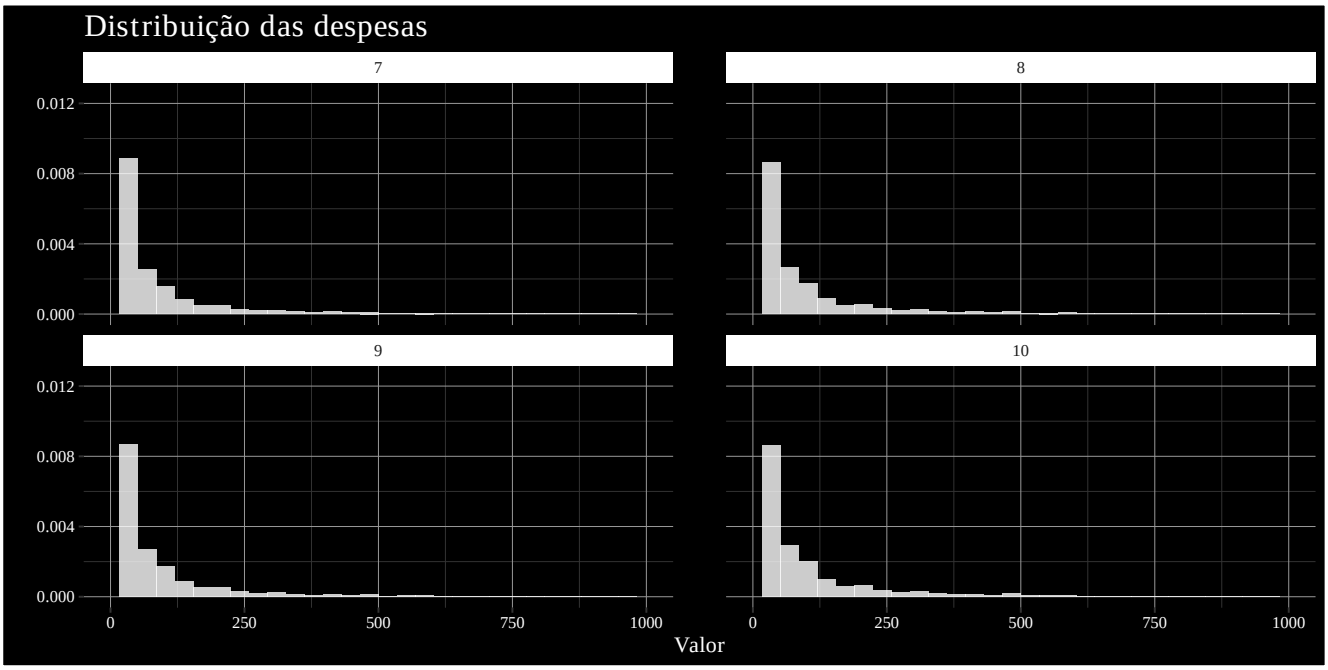
Para plotar o Histograma dos Valores gastos(Despesas) vamos limitar a variável 'Valor' em até 1000 reais. Tendo em vista que quase a totalidade dos dadados se concentram nesse intervalo.

```
Despesas %>%
  group_by(UsuarioId,
           dia = lubridate::day(DataDespesa),
           mes = lubridate::month(DataDespesa),
           ano = lubridate::year(DataDespesa)) %>%
    summarise(count = n(), valorSoma= sum(Valor)) %>%
    arrange(desc(valorSoma)) -> desp
```

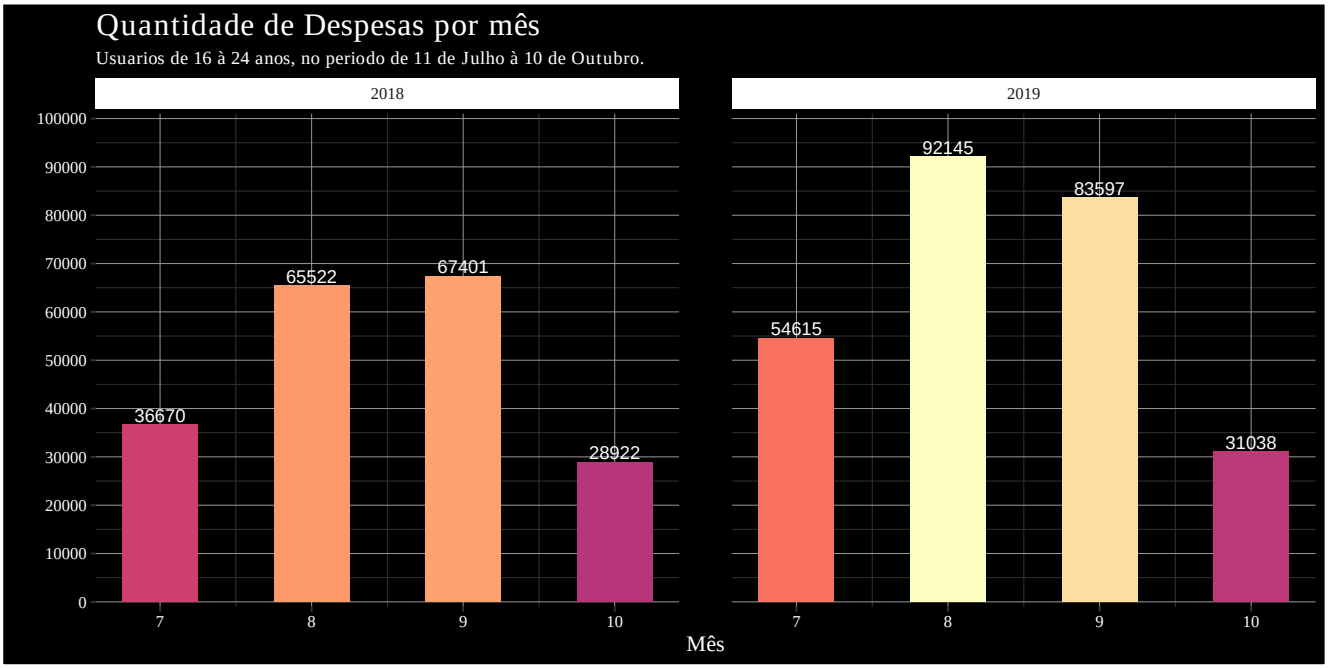
```
Despesas %>% dplyr::filter(Valor > 0 & Valor < 1000) %>%
    ggplot(aes(Valor,y=..density..))+
    geom_histogram(fill="white",alpha=0.8)+
    facet_grid(~ano )+temaMobills+
        labs(title="Distribuição das despesas")+
    scale_x_continuous(limits=c(0,1000))
```



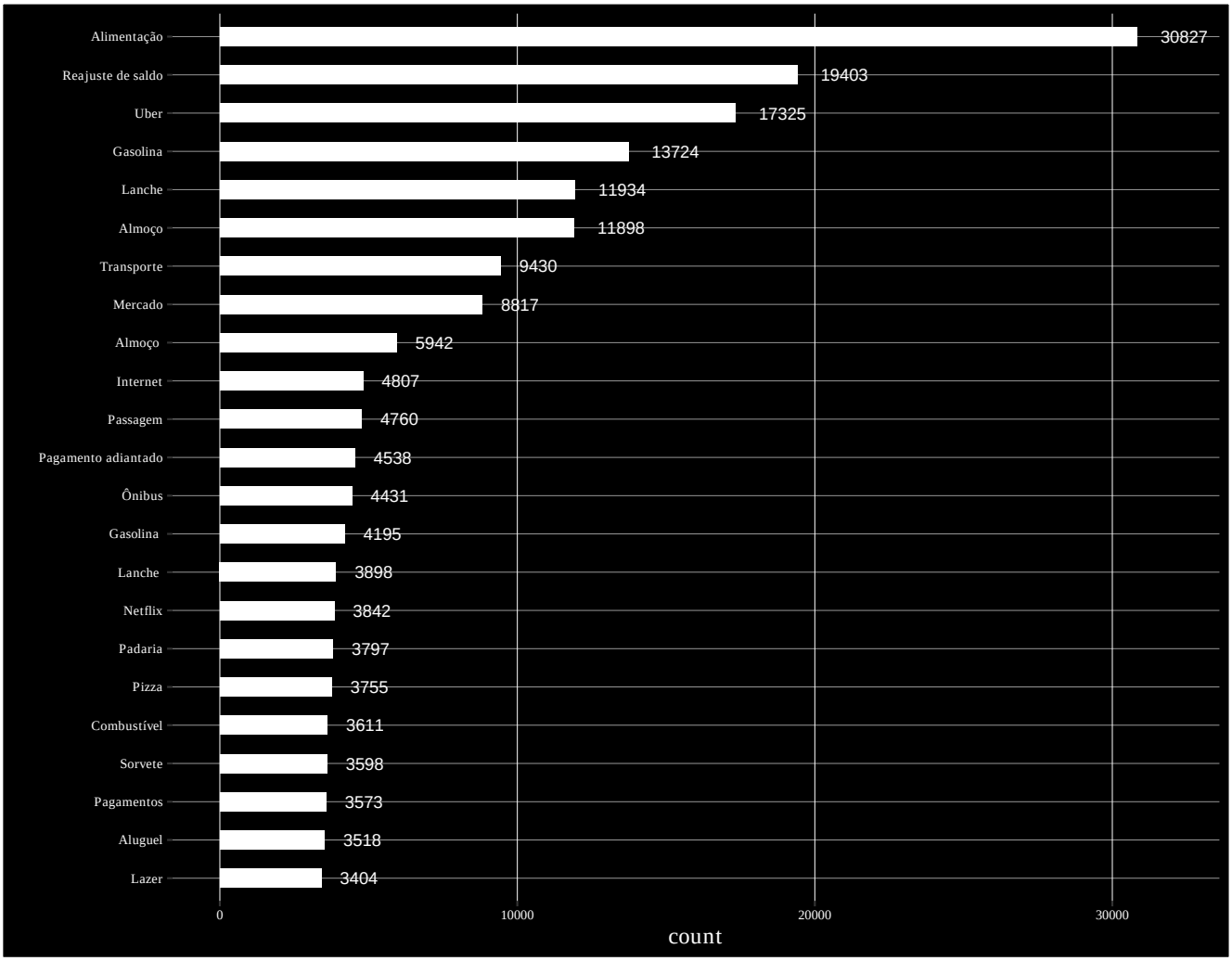

```
##histogram by month
Despesas %>% dplyr::filter(Valor > 0 & Valor < 1000) %>%
    ggplot(aes(Valor,y=..density..))+
    geom_histogram(fill="white",alpha=0.8)+
    facet_grid(~ano )+temaMobills+
        labs(title="Distribuição das despesas")+
    scale_x_continuous(limits=c(0,1000))+
  facet_wrap(~mes)
```



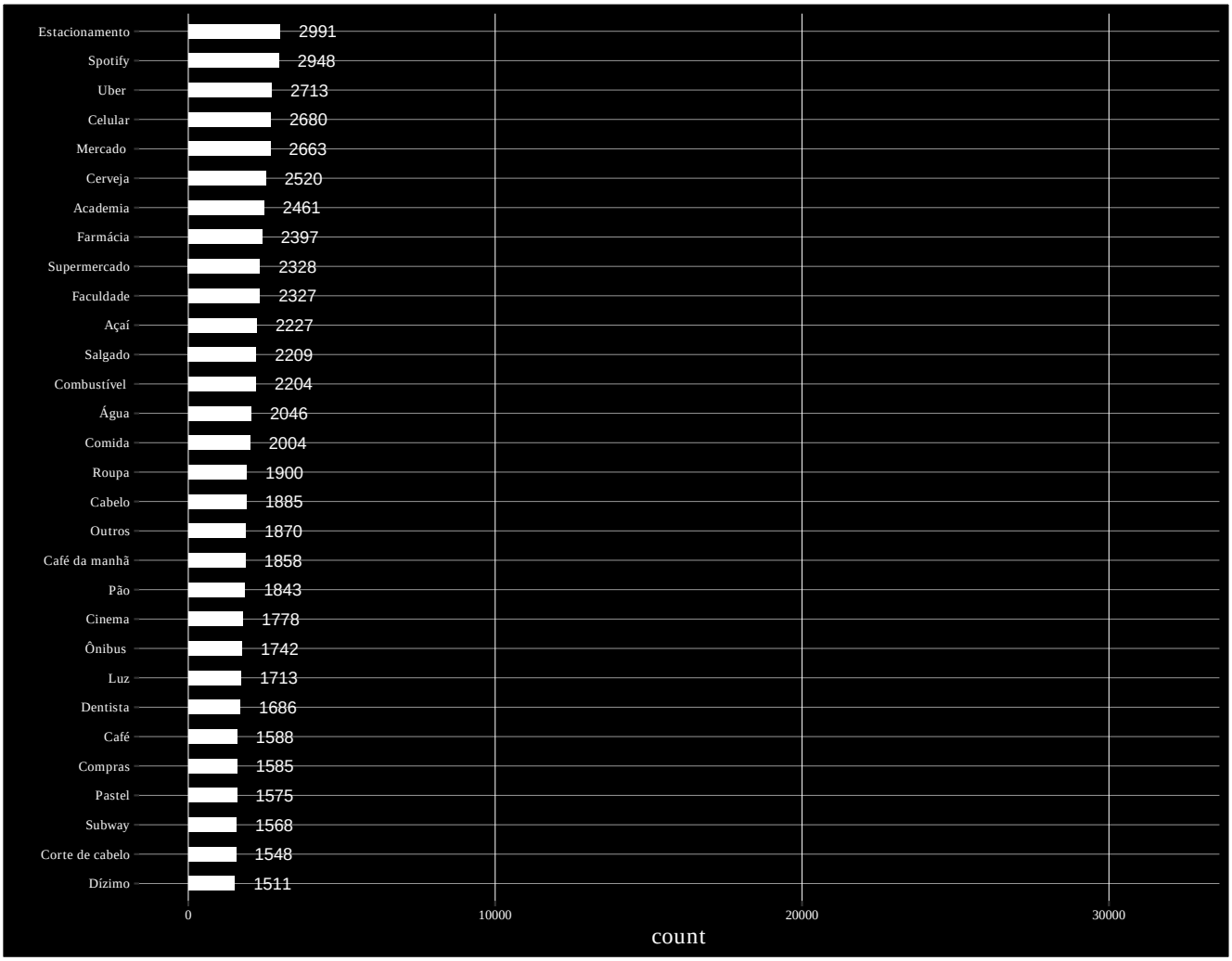

```
desp %>%
  group_by(mes,ano) %>%
  summarise(contagem=n()) %>%
  ggplot(aes(mes, contagem,labs=contagem))+
  geom_col(aes(fill=contagem),
           width = 0.5)+
  scale_fill_viridis(option="magma",begin=0.5)+
  labs(title="Quantidade de Despesas por mês",
       subtitle = "Usuarios de 16 à 24 anos, no periodo de 11 de Julho à 10 de Outubro.",
       x="Mês",
       y="Quantidade de despesas")+
  temaMobills+
  scale_y_continuous(limits = c(0,100000),
                     expand=c(0.01009,0.000000001),
                     breaks = seq(0,150000,10000))+
  geom_text(aes(label=contagem),
            size=3.5,
            colour="white",
            vjust=-0.2)+
  facet_grid(~ano)
```




```
Despesas %>%
    group_by(Descricao) %>%
    summarise(count = n(), valorSoma= sum(Valor)) %>%
    top_n(100000) %>% filter(count > 3000) %>% arrange(desc(count))%>%
  ggplot(aes(x=reorder(Descricao,count,max),count),labels=count)+
  geom_col(fill="white",width = 0.5)+
  coord_flip()+
  temaMobills+
  theme(axis.text = element_text(size=7),
        panel.grid.major.x =element_line(colour="white",linetype = 1),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(size=0.1))+
  geom_text(aes(label=count),colour="white",size=3,hjust=-0.5)+
  scale_y_continuous(limits=c(0,32000))
```




```
Despesas %>%
    group_by(Descricao) %>%
    summarise(count = n(), valorSoma= sum(Valor)) %>%
    top_n(100000) %>% filter(count <3000,count>1500) %>%
  arrange(desc(count))%>%
  ggplot(aes(x=reorder(Descricao,count,max),count),labels = count)+
  geom_col(fill="white",width = 0.5)+
  coord_flip()+
  temaMobills+
  theme(axis.text = element_text(size=7),
        panel.grid.major.x =element_line(colour="white",linetype = 1),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(size=0.1))+
  geom_text(aes(label=count),colour="white",size=3,hjust=-0.5)+
  scale_y_continuous(limits=c(0,32000))
```



Qual o tipo de despesa com o maior gasto total?

```
Despesas %>%
    dplyr::filter(Valor > 0 & Valor < 2000) %>%
    group_by(Descricao) %>%
    summarise(count = n(), valorSoma= sum(Valor)) %>%
    arrange(desc(valorSoma)) %>% top_n(20)

## # A tibble: 20 x 3
##    Descricao           count valorSoma
##    <chr>               <int>     <dbl>
##  1 Reajuste de saldo   19122  2043250.
##  2 Aluguel              3471  1735572.
##  3 Alimentação         30777  1058510.
##  4 Faculdade            2313   793314.
##  5 Gasolina            13724   659972.
##  6 Pagamento adiantado  2389   626031.
##  7 Carro                1369   618208.
##  8 Pagamentos           3542   543463.
##  9 "Aluguel "            994   465695.
## 10 Mercado              8786   411919.
## 11 Celular              2670   396889.
```

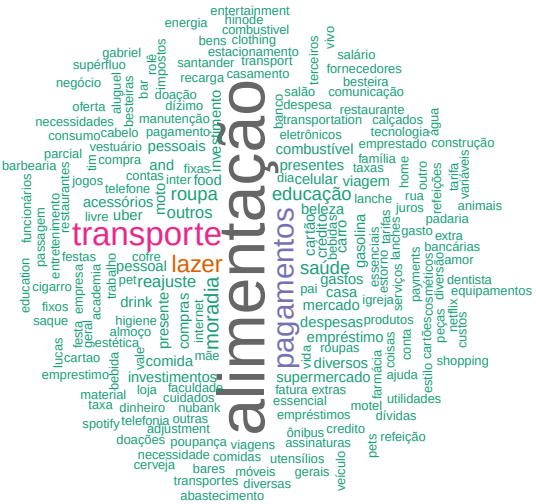
```
## 12 Internet             4804   376046.
## 13 Nubank               1012   341527.
## 14 Moradia              1389   331761.
## 15 "Faculdade "          834   274922.
## 16 Transporte           9393   270631.
## 17 Uber                17317   261770.
## 18 Moto                  974   239474.
## 19 Empréstimo            911   231605.
## 20 Almoço              11898   225157.
```

Agora iremos agrupar as depesas por categoria

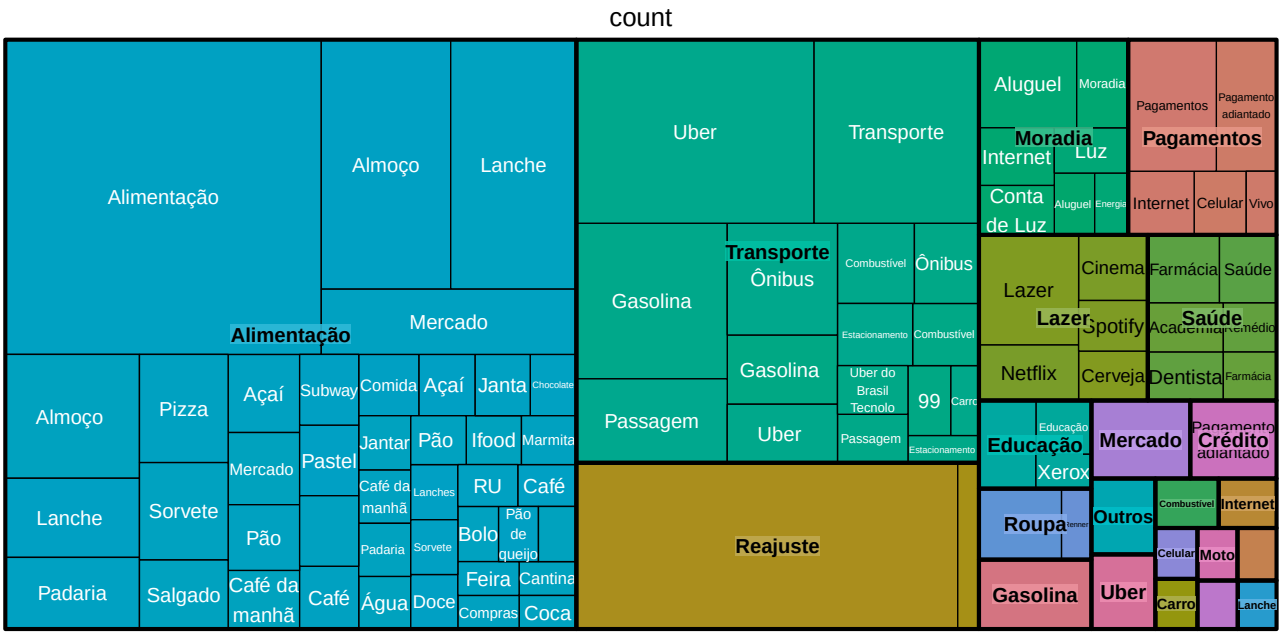
```
DespesasCat <- mutate(DespesasCat,
                      chave = paste0(DespesasCat$Id,DespesasCat$UsuarioId))
Despesas <- mutate(Despesas,
                      chave = paste0(Despesas$TipoDespesaId,Despesas$UsuarioId))
Despesas2 <- left_join(Despesas,DespesasCat,by=c('chave' = 'chave'))

Despesas2 %>% mutate(chaveUnica = paste0(Descricao, Nome, UsuarioId.x),
                     dia = lubridate::day(Despesas2$DataDespesa),
                     mes = lubridate::month(Despesas2$DataDespesa),
                     ano = lubridate::year(Despesas2$DataDespesa)) -> Despesas2

##Despesas2 %>% select(Descricao,
##Nome,
##TipoDespesaId,
##UsuarioId.x,
##UsuarioId.y,
##chaveUnica,
##mes) %>%
##distinct() %>% View()
##Despesas2[unique(Despesas2$chaveUnica), ]
##Despesas2[duplicated(Despesas2$chaveUnica), ]%>% View
##length(Despesas2$chaveUnica)


##Despesas2 %>% group_by(Descricao,Nome,mes) %>%
##summarise(contagem= n()) %>%
##top_n(100) %>% View()
```

```
text <- readLines("./texto.txt")
docs <- Corpus(VectorSource(text))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
tm_ma

## Error in eval(expr, envir, enclos):  objeto 'tm_ma' não encontrado
```

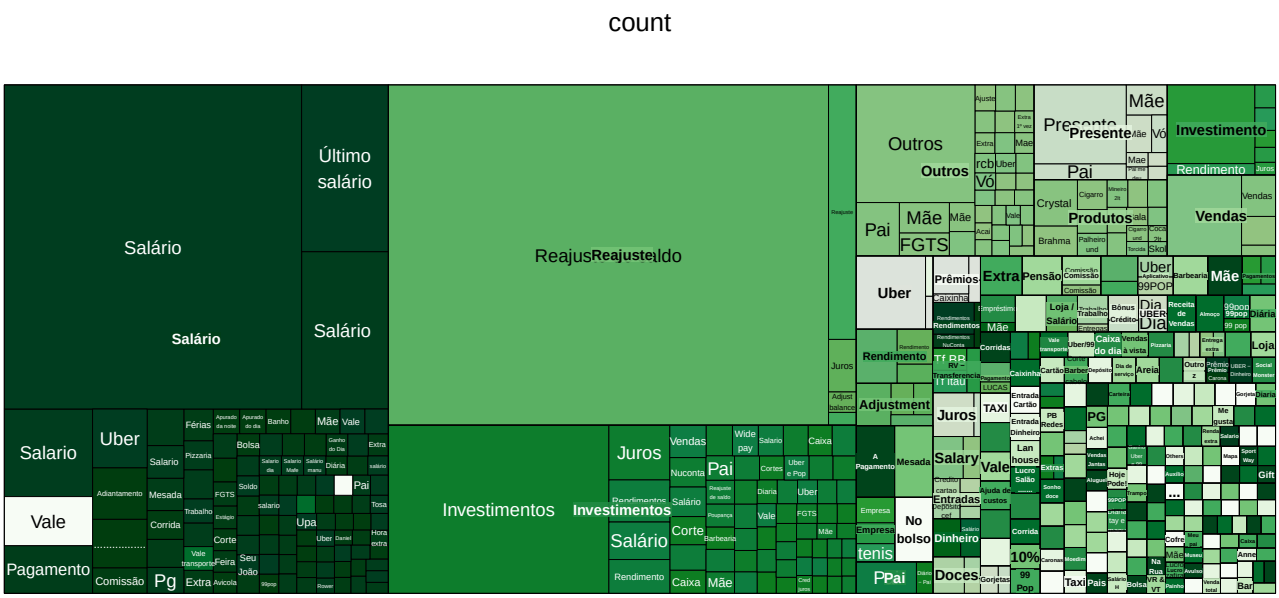
```
docs <- tm_map(docs, content_transformer(tolower))
# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("portuguese"))
# Remove your own stop word
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c("blabla1", "blabla2"))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
# Text stemming
# docs <- tm_map(docs, stemDocument)

dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d,20)

##                    word   freq
## alimentação alimentação 322355
## transporte   transporte 123890
## pagamentos   pagamentos  99637
## lazer             lazer  69323
## moradia         moradia  39244
## saúde             saúde  35075
## roupa             roupa  32306
## educação       educação  26014
## outros           outros  25733
## reajuste       reajuste  22964
## cartão           cartão  12876
## compras         compras  11744
## despesas       despesas  10086
## mercado         mercado  10069
## crédito         crédito   9873
## carro             carro   9846
## casa               casa   9595
## gastos           gastos   9481
## beleza           beleza   8631
## presente       presente   7593

set.seed(1234)
wordcloud(words = d$word, freq = d$freq,scale=c(3,0.6), min.freq = 500,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```




```
# Build Dataset

Despesas2 %>% select(Nome,Descricao) %>%
  group_by(Nome,Descricao) %>%
  summarise(count=n()) %>% arrange(desc(count)) %>% top_n(50) %>% head(100) %>%
  treemap(index=c("Nome","Descricao"),
            vSize="count",
            vColor = "Nome",
            type="index",range=c(0,10000)
            )
```



2 Análise das Receitas




```
textr <- readLines("./textoReceitas.txt")
docsr <- Corpus(VectorSource(textr))
docsr <- tm_map(docsr, toSpace, "/")
docsr <- tm_map(docsr, toSpace, "@")
docsr <- tm_map(docsr, toSpace, "\\|")

docsr <- tm_map(docsr, content_transformer(tolower))
# Remove numbers
docsr <- tm_map(docsr, removeNumbers)
# Remove english common stopwords
docsr <- tm_map(docsr, removeWords, stopwords("portuguese"))
# Remove your own stop word
# specify your stopwords as a character vector
##docsr <- tm_map(docsr, removeWords, c("blabla1", "blabla2"))
# Remove punctuations
docsr <- tm_map(docsr, removePunctuation)
# Eliminate extra white spaces
docsr <- tm_map(docsr, stripWhitespace)
# Text stemming
# docs <- tm_map(docs, stemDocument)

dtmr <- TermDocumentMatrix(docsr)
mr <- as.matrix(dtmr)
vr <- sort(rowSums(mr),decreasing=TRUE)
dr <- data.frame(word = names(vr),freq=vr)
head(dr,10)

##                       word  freq
## salário             salário 61896
```

```
## investimentos investimentos 33263
## reajuste            reajuste 28335
## outros                outros 27834
## presente            presente  6928
## vendas                vendas  6553
## investimento   investimento  5618
## empréstimo       empréstimo  4036
## cartão                cartão  3767
## pagamento          pagamento  3480

set.seed(1234)
wordcloud(words = dr$word, freq = dr$freq,scale=c(3,0.4), min.freq = 200,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```