Mobills

Relatório de Análise Descritiva

FORTALEZA
2019

1 Análise das despesas

A seguir temos o TOP 100 das Despesas que mais aparecem no dataset:
As medidas de resumo mostram que claramente há valores inválidos e outliers nos dados.

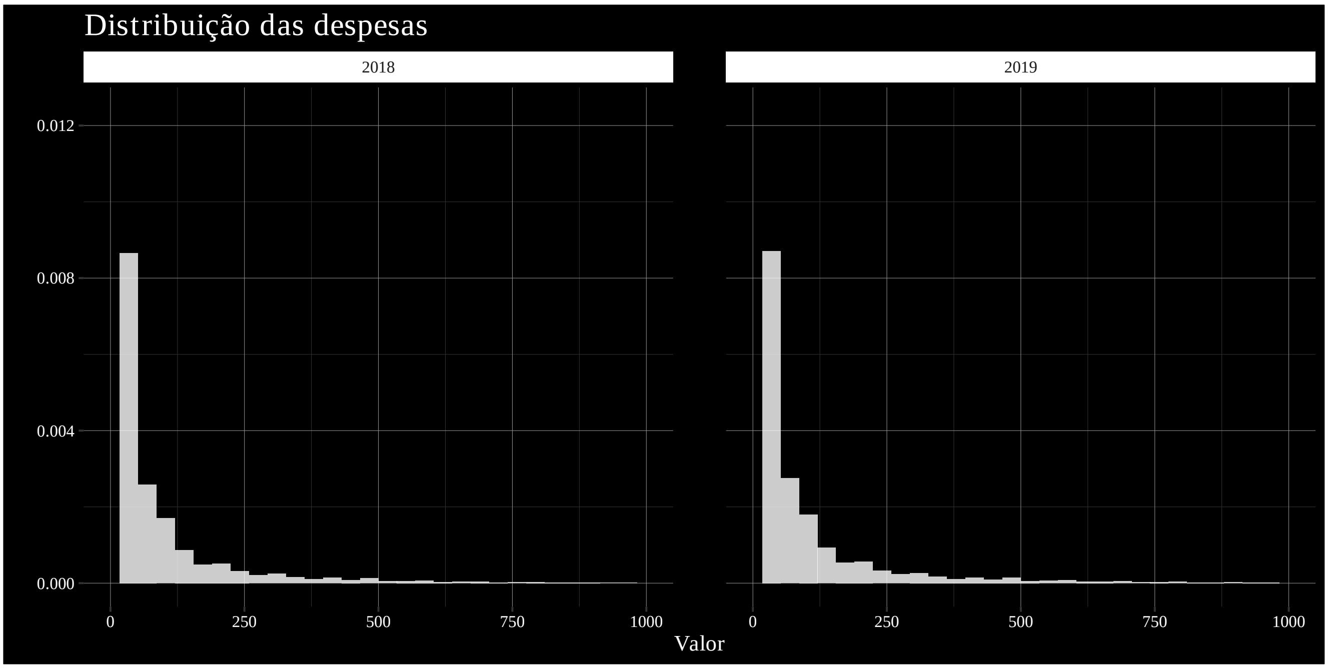
```
Despesas %>%
  select(ano,Valor)%>%
  split(.$ano) %>%
  map(summary)

## $`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. T endo em vista que quase a totalidade dos dadados se concentram nesse intervalo.

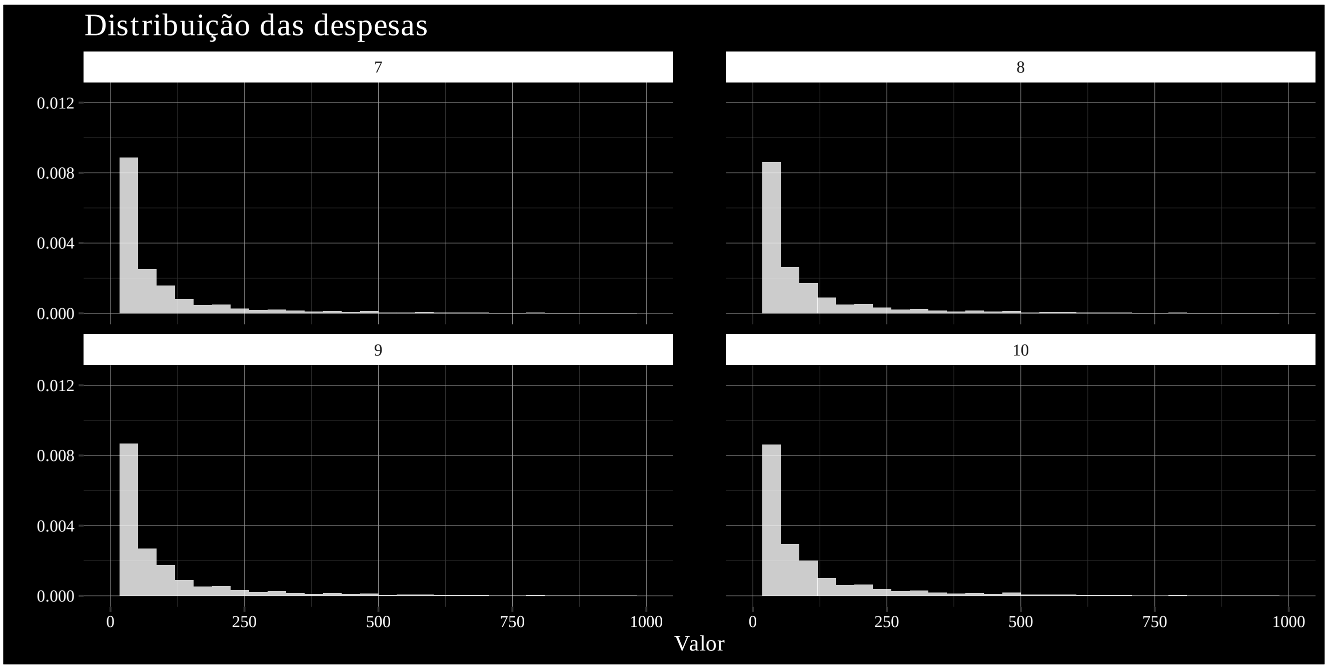
```
Despesas %>%
  group_by(UsuarioId,
           dia = lubridate::day(DataDespesa),
           mes = lubridate::month(DataDespesa),
           ano = lubridate::year(DataDespesa)) %>%
    summarise(count = n(), valorSoma= sum(Valor)) %>%
    arrange(desc(valorSoma)) -> desp

head(desp,20) %>% kable(format = "latex")
```

UsuarioId	dia	mes	ano	count	valorSoma
2e485e2a-5371-4608-9570-020aedff3af8	9	8	2019	4	9089000000000030.0
2e485e2a-5371-4608-9570-020aedff3af8	16	7	2019	14	305168230035490.4
fa5c03e3-f075-4833-a311-1f88db84d713	6	9	2019	6	8022422222222.0
da0ca00e-5951-4370-b9a1-fb1336f14179	22	8	2019	18	18873919.2
898035e2-ee5a-492b-aa28-1385a0e2f8f8	10	8	2019	3	10000011.9
32727396-fe7d-49c3-a43c-e084f82c3d0d	14	9	2018	5	9300000.0
32727396-fe7d-49c3-a43c-e084f82c3d0d	15	9	2018	29	4144579.5
da0ca00e-5951-4370-b9a1-fb1336f14179	21	8	2019	8	3110644.9
72075617-5b61-4bd9-b853-0c2883b71bc7	4	9	2018	5	1106000.0
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	5	8	2018	9	890054.1
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	5	9	2018	9	890054.1
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	5	10	2018	9	890054.1
72075617-5b61-4bd9-b853-0c2883b71bc7	6	8	2018	5	859000.0
72075617-5b61-4bd9-b853-0c2883b71bc7	2	10	2018	4	812000.0
72075617-5b61-4bd9-b853-0c2883b71bc7	3	9	2019	7	730500.0
da0ca00e-5951-4370-b9a1-fb1336f14179	23	8	2019	1	537780.0
72075617-5b61-4bd9-b853-0c2883b71bc7	17	9	2018	5	519000.0
c61685d4-460d-48a5-baaf-33736e76e428	21	9	2018	2	493225.1
72075617-5b61-4bd9-b853-0c2883b71bc7	3	10	2019	6	491000.0
72075617-5b61-4bd9-b853-0c2883b71bc7	26	9	2018	2	460000.0

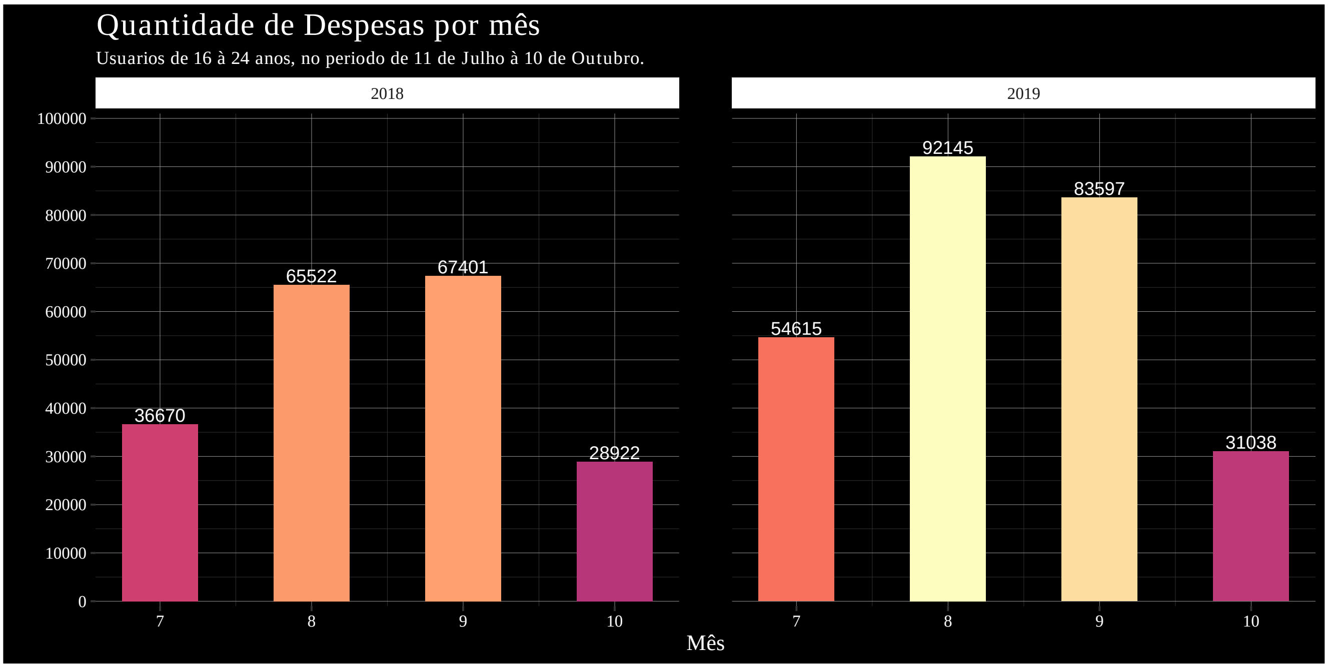
```r
Despesas %>% dplyr::filter(Valor > 0 & Valor < 1000) %>%
    ggplot(aes(Valor,y=..density..))+
    geom_histogram(fill="white",alpha=0.8)+
    facet_grid(~ano )+temaMobills+
        labs(title="Distribuição das despesas")+
    scale_x_continuous(limits=c(0,1000))
```



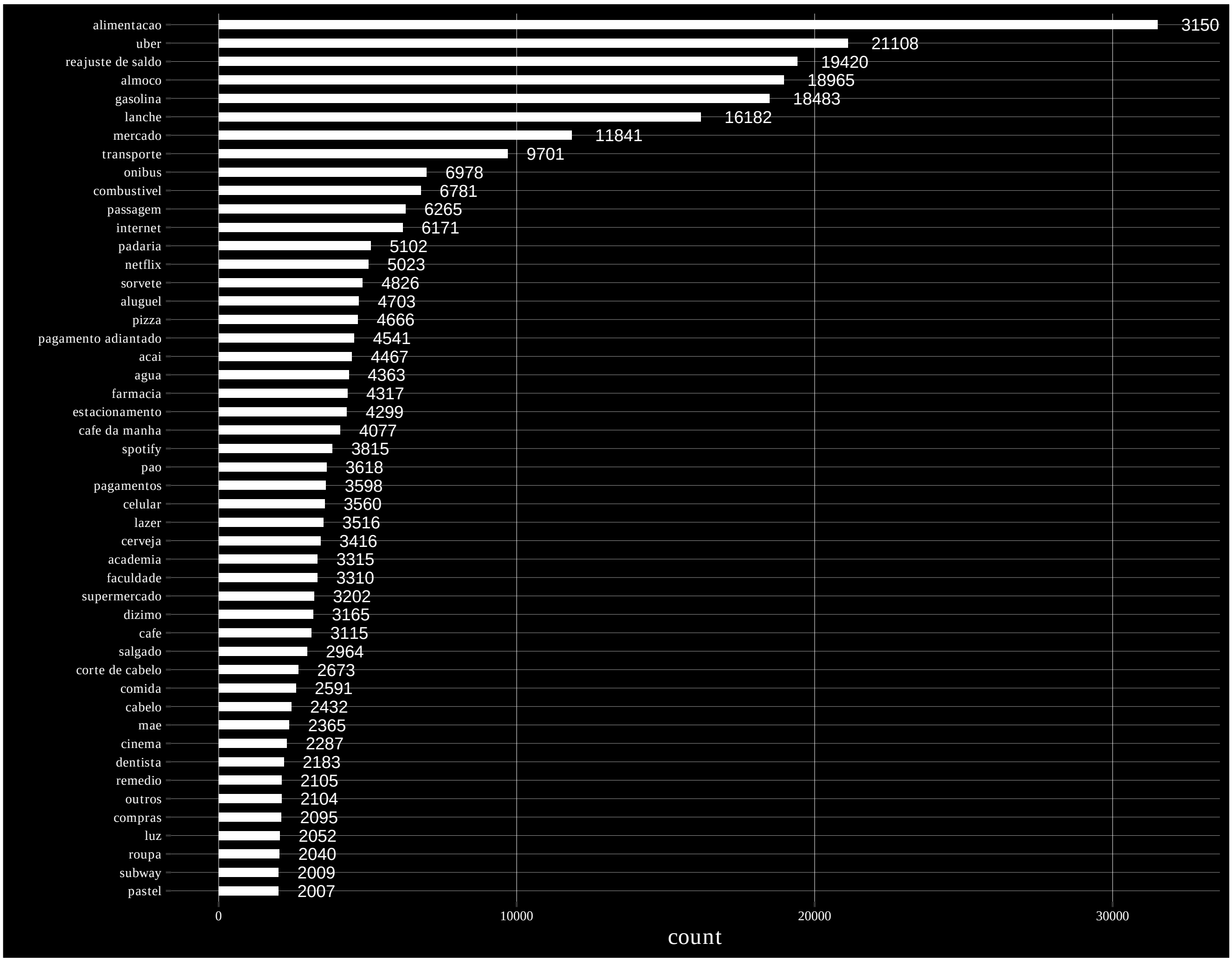
```r
##histogram by month
Despesas %>% dplyr::filter(Valor > 0 & Valor < 1000) %>%
    ggplot(aes(Valor,y=..density..))+
    geom_histogram(fill="white",alpha=0.8)+
    facet_grid(~ano )+temaMobills+
        labs(title="Distribuição das despesas")+
    scale_x_continuous(limits=c(0,1000))+
  facet_wrap(~mes)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



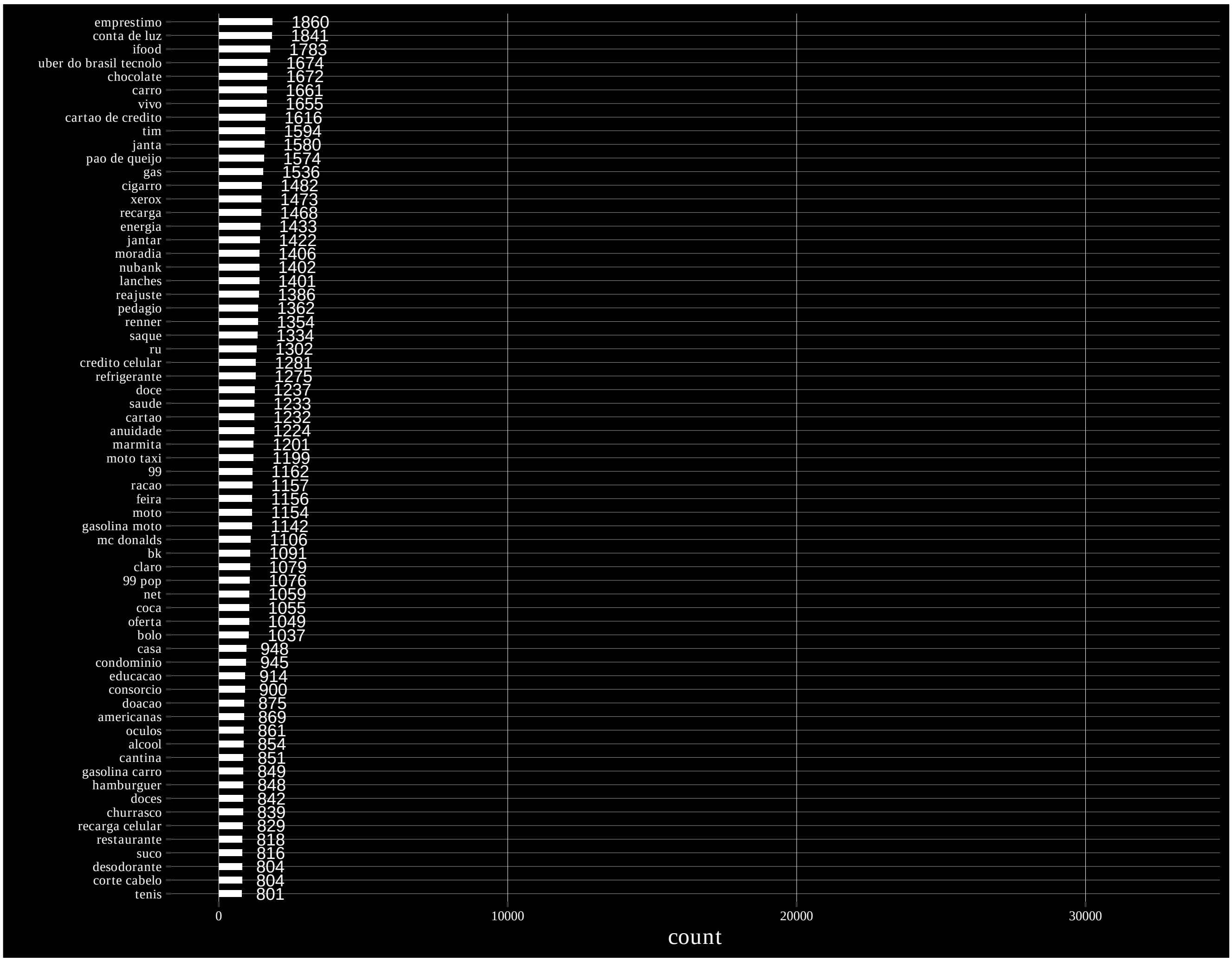

```r
desp %>%
  group_by(mes,ano) %>%
  summarise(contagem=n()) %>%
  ggplot(aes(mes, contagem,labs=contagem))+
  geom_col(aes(fill=contagem),
           width = 0.5)+
  scale_fill_viridis(option="magma",begin=0.5)+
  labs(title="Quantidade de Despesas por mês",
       subtitle = "Usuarios de 16 à 24 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 > 2000) %>% 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 < 2000,
         count > 800) %>%
  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(0,33000))
```



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 aluguel              4636  2284457.
##  2 reajuste de saldo   19139  2043957.
##  3 faculdade            3285  1119109.
##  4 alimentacao         31458  1087975.
##  5 gasolina            18479   878696.
##  6 carro                1635   737825.
##  7 pagamento adiantado  2389   626031.
##  8 mercado             11810   550892.
##  9 celular              3547   546548
## 10 pagamentos           3567   546360.
## 11 cartao de credito    1595   543311.
```

```
## 12 internet              6167   479865.
## 13 emprestimo            1843   479166.
## 14 nubank                1382   471088.
## 15 dizimo                3161   378172.
## 16 combustivel           6769   375848.
## 17 almoco               18965   365202.
## 18 cartao                1214   361481.
## 19 moradia               1399   334209.
## 20 uber                 21098   318556.
```

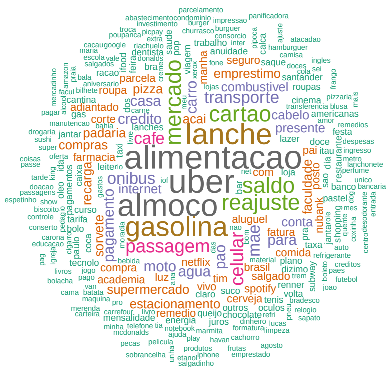
Agora iremos agrupar as depesas por categoria

```
DespesasCat$Nome <- gsub(pattern = "\"",replacement = "", DespesasCat$Nome)
DespesasCat$Nome <- gsub(pattern = "|" ,replacement = "", DespesasCat$Nome)
DespesasCat$Nome <- trim(DespesasCat$Nome)
DespesasCat$Nome <- tolower(DespesasCat$Nome)
DespesasCat$Nome <- rm_accent(DespesasCat$Nome)


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()
```

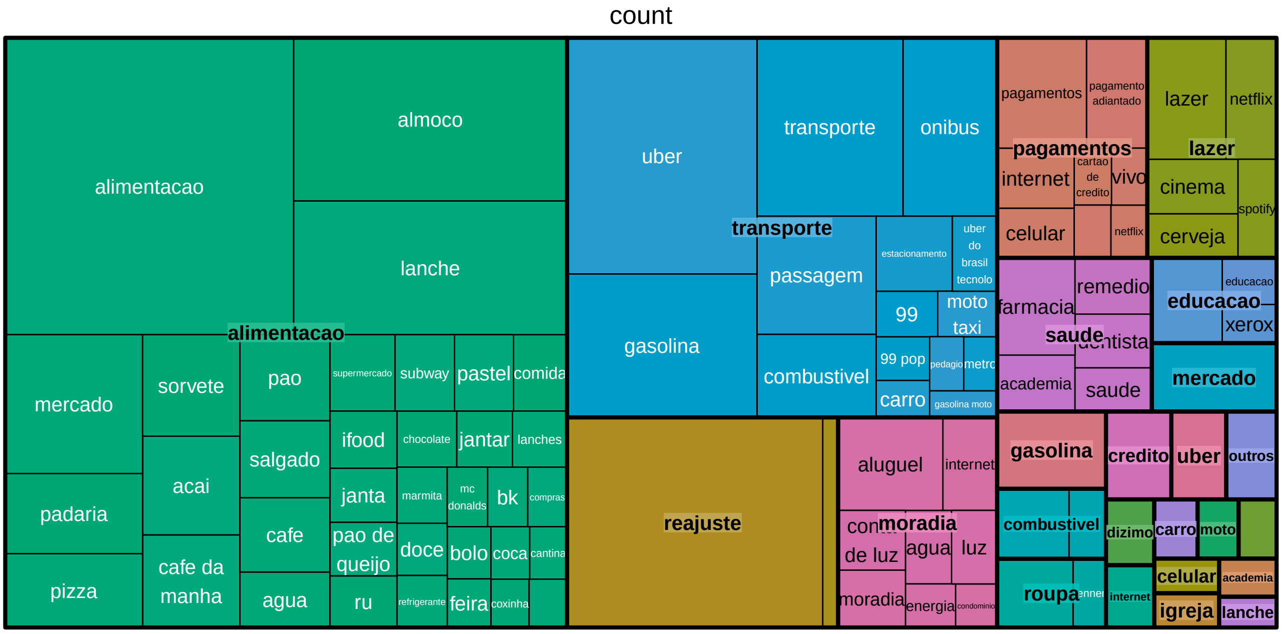
```
##text <- Despesas2$Descricao %>% paste(collapse = " ")
##write(text,"~/Mobills1stReport/data/textDespesasDesc.txt")
textDespesasDesc <- readLines("~/Mobills1stReport/data/textDespesasDesc.txt")
docsDespDesc <- Corpus(VectorSource(textDespesasDesc))
docsDespDesc <- tm_map(docsDespDesc, toSpace, "/")
docsDespDesc <- tm_map(docsDespDesc, toSpace, "@")
docsDespDesc <- tm_map(docsDespDesc, toSpace, "\\|")

##docs <- tm_map(docs, content_transformer(tolower))
# Remove numbers
docsDespDesc <- tm_map(docsDespDesc, 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
docsDespDesc <- tm_map(docsDespDesc, removePunctuation)
# Eliminate extra white spaces
#docs <- tm_map(docs, stripWhitespace)
# Text stemming
# docs <- tm_map(docs, stemDocument)
dtm <- TermDocumentMatrix(docsDespDesc)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d,30)
```

```
##                   word  freq
## uber              uber 34272
## alimentacao alimentacao 33713
## almoco          almoco 30903
## lanche          lanche 29799
## gasolina      gasolina 27046
## reajuste      reajuste 21069
## mercado        mercado 20423
## saldo            saldo 19890
## cartao          cartao 19392
## passagem      passagem 16241
## celular        celular 15789
## cafe              cafe 13168
## onibus          onibus 12780
## mae                mae 11980
## transporte  transporte 11736
## casa              casa 11088
## credito        credito 10997
## pao                pao 10837
## moto              moto 10673
## agua              agua  9791
## para              para  9678
```

```
## carro               carro  9669
## pagamento       pagamento  9599
## combustivel combustivel  9274
## internet         internet  9194
## conta               conta  8929
## presente         presente  8719
## cabelo             cabelo  8708
## emprestimo   emprestimo  8512
## pizza               pizza  8238

set.seed(1234)
wordcloud(words = d$word, freq = d$freq,scale=c(1.5,0.3), min.freq = 1000,
          max.words=1000, random.order=FALSE, rot.per=0.25,
          colors=brewer.pal(8, "Dark2"))
```



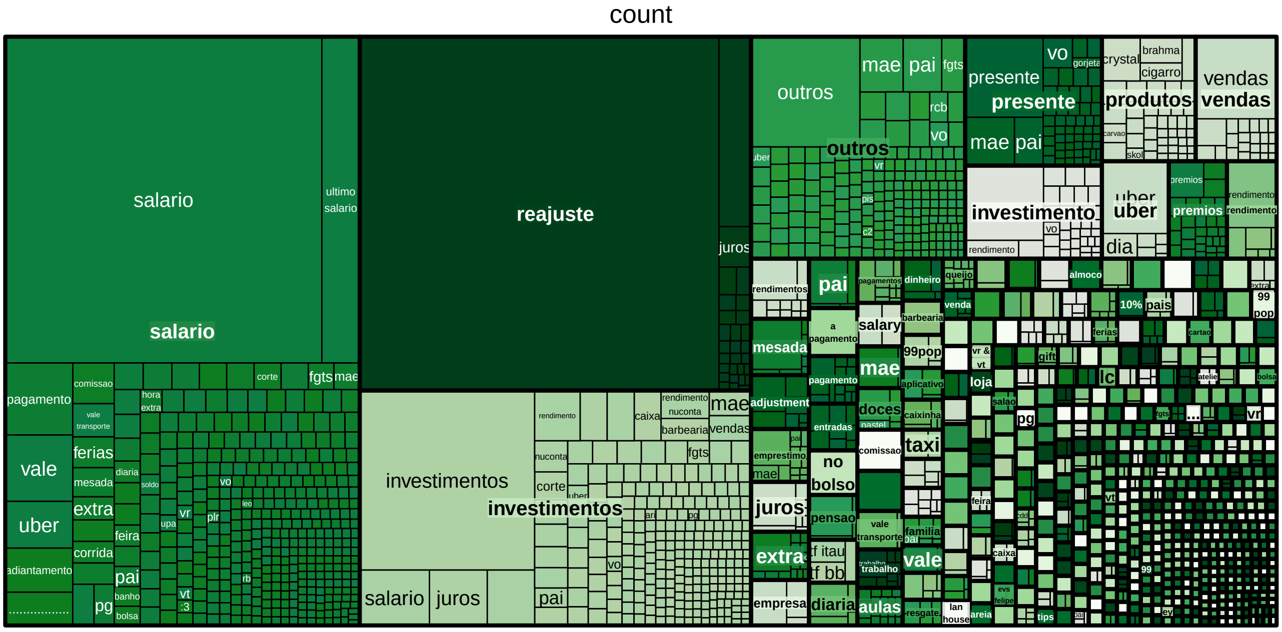

```
# Build Dataset

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

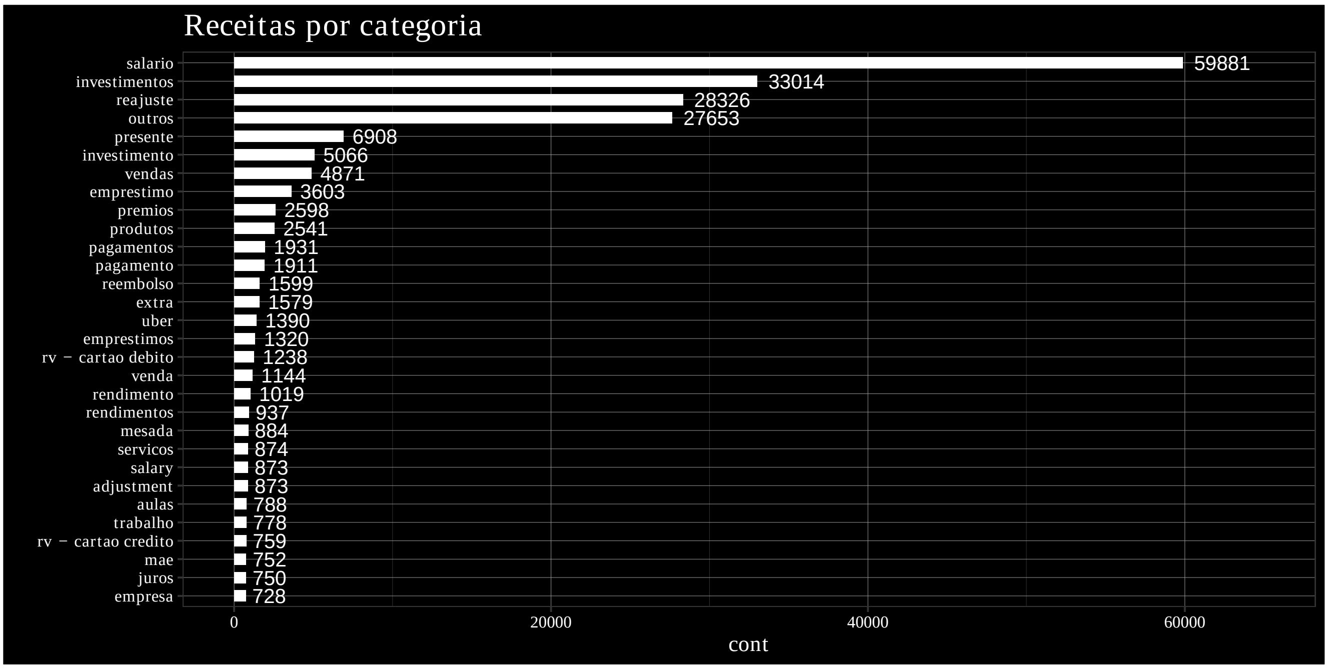
```
Receitas %>%
  group_by(UsuarioId,
           dia = lubridate::day(DataReceita),
           mes = lubridate::month(DataReceita),
           ano = lubridate::year(DataReceita)) %>%
    summarise(count = n(), valorSoma= sum(Valor)) %>%
    arrange(desc(valorSoma)) -> Rec

head(Rec,20) %>% kable(format = "latex")
```

UsuarioId	dia	mes	ano	count	valorSoma
fa5c03e3-f075-4833-a311-1f88db84d713	6	9	2019	9	8021470666666
fb7c9c59-df55-48e3-9871-7b67d3716f7f	2	8	2018	1	1051999990
da4b4ed3-168d-46f6-84ef-9c0af484ef6d	9	9	2019	1	600000000
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	15	8	2018	2	119365518
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	15	8	2019	2	119365518
da0ca00e-5951-4370-b9a1-fb1336f14179	18	8	2019	1	90000000
da0ca00e-5951-4370-b9a1-fb1336f14179	22	8	2019	1	85458556
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	8	9	2018	1	31838520
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	8	9	2019	1	31838520
f0bb4720-5dac-4095-bd06-63df9b2e7e0c	27	9	2018	1	23082018
25a966e2-2ada-489e-925b-48feee87af17	10	8	2019	2	10000250
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	28	7	2018	1	4968210
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	28	7	2019	1	4968210
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	28	8	2018	1	4968210
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	28	8	2019	1	4968210
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	28	9	2018	1	4968210
b6a87d2d-836c-4ea0-a90d-3445d0dde5ba	28	9	2019	1	4968210
0d7a40bf-13b9-4f46-8f55-f0e6c96e3f94	24	7	2019	1	2500000
fa5c03e3-f075-4833-a311-1f88db84d713	4	10	2019	1	2222222
6a82d95b-8714-4b30-a682-d0752c4ba07b	6	9	2018	1	2000000

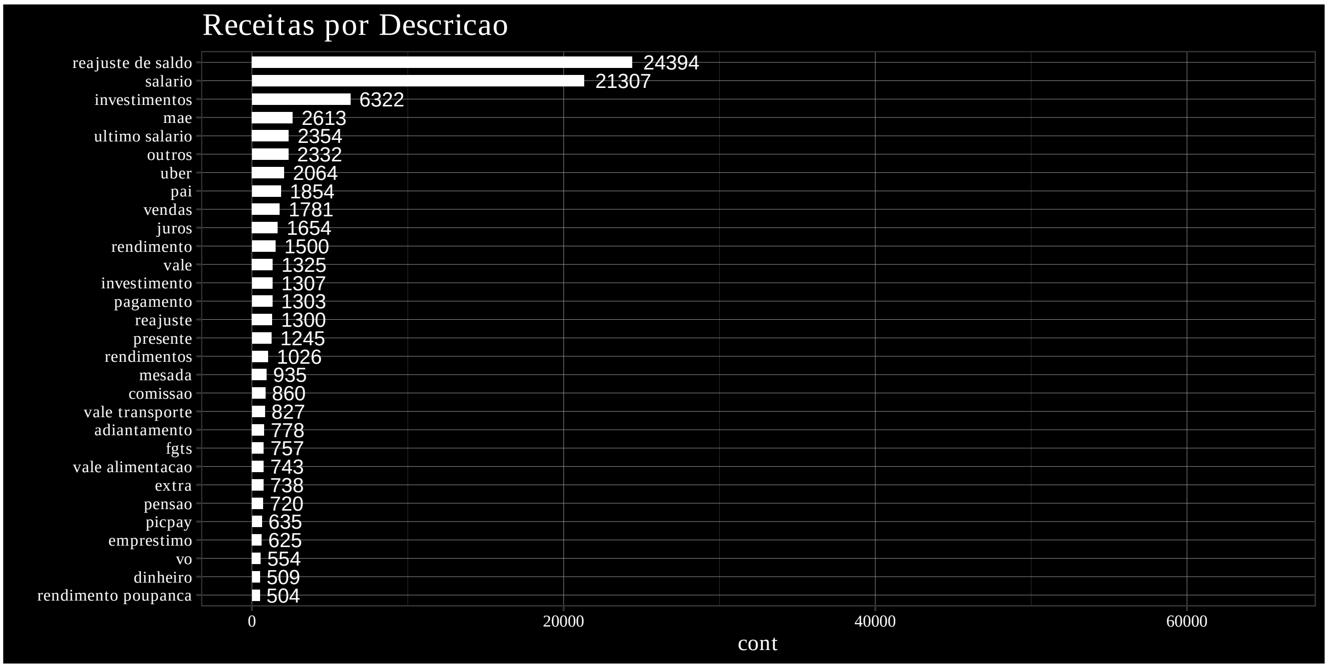



```
Receitas2 %>% group_by(Nome) %>%
  summarise(cont=n()) %>%
  arrange(desc(cont)) %>%
  top_n(30) %>%
  ggplot(aes(reorder(Nome,cont,max),labs=cont,cont))+
  geom_col(fill="white",width=0.6)+
  theme_bw()+
  coord_flip()+
  temaMobills+
```

```
  labs(title="Receitas por categoria")+
  geom_text(aes(label=cont),colour="white",hjust=-0.2)+
  scale_y_continuous(limits=c(0,65000))
```




```
Receitas2 %>% group_by(Descricao) %>%
  summarise(cont=n()) %>%
  arrange(desc(cont)) %>%
  top_n(30) %>%
  ggplot(aes(reorder(Descricao,cont,max),labs=cont,cont))+
  geom_col(fill="white",width=0.6)+
  theme_bw()+
  coord_flip()+
  temaMobills+
  labs(title="Receitas por Descricao")+
  geom_text(aes(label=cont),colour="white",hjust=-0.2)+
  scale_y_continuous(limits=c(0,65000))
```




```r
textr <- readLines("~/Mobills1stReport/data/textoReceitasNome.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
## salario            salario 62289
## investimentos investimentos 33242
## reajuste          reajuste 28335
## outros              outros 27826
```

```
## presente           presente  6928
## vendas               vendas  6548
## investimento   investimento  5618
## emprestimo       emprestimo  4655
## cartao               cartao  3820
## pagamento         pagamento  3473
```