PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Itanimá Baroni

O REGIME DE TRIBUTAÇÃO NO CICLO DE VIDA DAS EMPRESAS

Araraquara
2020

Itanimá Baroni

O REGIME DE TRIBUTAÇÃO NO CICLO DE VIDA DAS EMPRESAS

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Araraquara
2020

SUMÁRIO

1. Introdução

1.1. Contextualização

Há muito temos notícias de que o ciclo de vida das empresas não consegue superar os cinco anos iniciais. O chamado “Tempo de Sobrevivência” é extremamente curto, agravando o risco da abertura de um novo negócio.

Contudo, sob a ótica das políticas econômicas, o estímulo ao empreendedorismo está na pauta de todos os discursos nas últimas décadas, o que a princípio parece um contrassenso.

O presente trabalho pretende avaliar o ciclo de vida das empresas com data de abertura de 2005 a 2020, possibilitando a verificação do comportamento das mesmas nos últimos 15 anos.

Considerando a diversidade dos regimes tributários ora vigentes no Brasil, separamos nossa base de análise em três regimes bastante distintos: O Micro empreendedor Individual – MEI; O Simples Nacional e o Lucro Real/Presumido.

As empresas optantes pelo Lucro Presumido e pelo Lucro Real foram agrupadas, pois elas diferem muito mais em relação aos dois outros grupos, do que propriamente entre si.

Utilizaremos a base de dados mantida pela Receita Federal do Brasil. denominada Cadastro Nacional de Pessoa Jurídica - CNPJ, onde estão presentes os dados históricos de todas as empresas registradas no Brasil, ativas ou não.

1.2. O problema proposto

Pretende-se com este trabalho avaliar a taxa de sobrevivência das empresas optantes pelo Simples Nacional, optantes pelo MEI - Microempreendedores Individuais, e optantes pelo Lucro Presumido/Real.

Para uma melhor visualização dos objetivos utilizaremos a técnica dos 5-Ws:

- (Why?): A perspectiva de se montar um negócio traz consigo o risco deste não prosperar. A oportunidade de conhecer melhor este risco possibilita ao empreendedor elaborar uma estratégia prévia, na busca de melhores resultados.

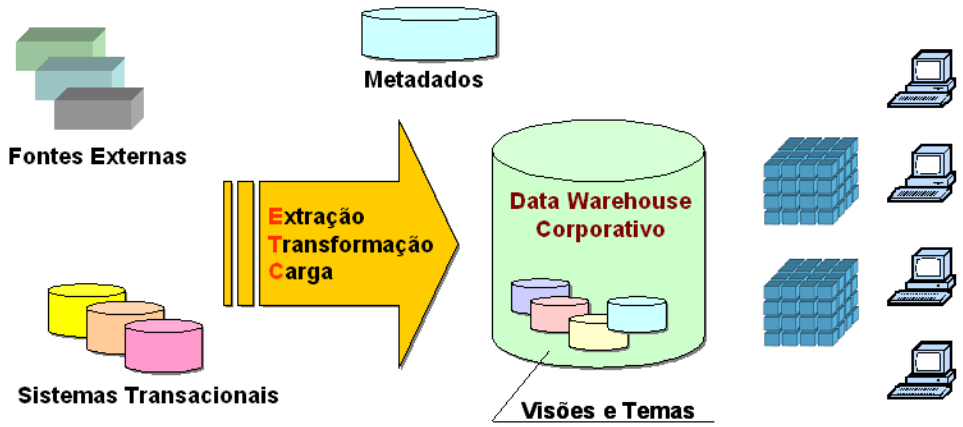
- (Who?): Os dados utilizados no presente trabalho foram obtidos de uma base governamental – Cadastro Nacional das Pessoas Jurídicas - CNPJ.
- (What?): Iremos analisar uma base cadastral das Pessoas Jurídicas no Brasil, onde estão contempladas todas as empresas ativas, ou não.
- (Where?): Concretizada as técnicas a serem utilizadas estas poderão ser disponibilizadas para utilização independente do local, considerando a disponibilidade das bases de consulta.
- (When?): Os dados representam as empresas inscritas no Cadastro do CNPJ a partir do ano de 2005, até a presente data.

O período utilizado na Análise Exploratória e treinamento dos modelos correspondem aos anos de 2005 a 2019. Para os dados de teste dos modelos treinados, foi utilizado o ano de 2020.

2. Coleta de Dados

Os dados que serão utilizados no presente trabalho foram obtidos diretamente do Cadastro Nacional da Pessoa Jurídica – CNPJ, extraídos do Data Warehouse, orientado por assunto, denominado DW, cuja administração cabe à Receita Federal do Brasil (RFB).

O processo de Data Wahehousing na RFB envolve etapas bem definidas: A) Extração das informações, das bases de dados transacionais e das bases de outros órgãos ou empresas; B) Transformação, com a limpeza e integração das informações; C) A carga numa base única, integrada e estruturada permitindo que as consultas necessárias sejam efetuadas pelos usuários das diversas áreas de negócio da RFB.



A partir desta base foram extraídas três bases de consulta: Uma contendo os dados das empresas optantes pelo Lucro Real/Presumido, ou seja, não optantes pelo Simples Nacional, e não optantes do MEI; Uma segunda contendo empresas optantes pelo Simples Nacional, e não optantes do MEI, e por fim uma terceira contendo as empresas optantes pelo MEI.

Cabe observar que não existe a possibilidade de uma mesma empresa fazer parte de mais de uma base.

Enfatizamos, também, que as consultas trazem dados dos estabelecimentos, ou seja, uma empresa poderá ter dois ou mais estabelecimentos, e os eventos de abertura e fechamento são individualizados por estabelecimento.

O DataSet denominado “LR”, conterá os dados empresas optantes pelo Lucro Real/Presumido. O relatório da extração dos dados do Data Warehouse está descrito abaixo.

Nome do relatório:Consulta CNPJs Empresas Lucro Real/Presumido
Proprietário:ITANIMA BARONI
Caminho do relatório:CNPJ>Meus relatórios>Consulta CNPJs
Modificado:12/09/20 14:58:53

Hora inicial:12/09/20 15:18:23
Tempo de término:12/09/20 15:18:31
Descrição do Relatório:
Filtro do relatório:
({Empresa - 1 Ano Abertura Atual} (Nome) >= "2005") E ({Empresa - Ind. Ativo SIMEI Atual} = Não:N) E ({Empresa - Ind. Op. Simples Nacional CNPJ Atual} = Não)

Template:
Estab. - 3 UF Atual
Unidade da Federação onde se localiza o endereço do estabelecimento.

Estab. - 1 Ano Abertura Atual
Ano da data da Abertura Atual
Estab. - 1 Ano Baixa Cad. Atual
Ano da data da Baixa Cad. atual
Estab. - Sit. Cad. Atual
Situação cadastral do estabelecimento
Estab. - Sit. Cad. Motivo Atual
Motivo da situação cadastral da estabelecimento.
Métrica:
Qtde Estabelecimento Atual
Quantidade de Estabelecimentos.
Count({Estab. - CNPJ Atual})
{~+}

Instrução SQL:

```
select      a16.SG_UNIDADE_FEDERACAO  SG_UNIDADE_FEDERACAO,
  a17.ANO  ANO,
  a112.NB_ANO  NB_ANO,
  a14.ANO  ANO0,
  a19.NB_ANO  NB_ANO0,
  a11.NR_CNPJ_ESTA_SCAD_AT  NR_SIT_CADASTRAL,
  a110.NM_SIT_CADASTRAL  NM_SIT_CADASTRAL,
  a11.NR_CNPJ_ESTA_SCAD_MOT_AT  NR_MOT_SIT_CADASTRAL,
  a111.NM_MOT_SIT_CADASTRAL  NM_MOT_SIT_CADASTRAL,
  count(distinct a11.B_CD_CNPJ_ESTA)  WJXBFS1
from      WD_CNPJ_ESTA  a11
  join      WD_DT_DIAS    a12
    on      (a11.NR_CNPJ_EMPA_ABER_DIA_AT = a12.DT_DIA)
  join      WD_PJ_ESTABELECIMENTOS      a13
    on      (a11.NR_CNPJ_ESTA = a13.NR_ESTABELECIMENTO)
  join      WD_DT_DIAS    a14
    on      (a11.NR_CNPJ_ESTA_BCAD_DIA_AT = a14.DT_DIA)
  join      WD_LC_MUNICIPIOS      a15
    on      (a11.NR_CNPJ_ESTA_LC_MC_AT = a15.NR_MUNICIPIO)
  join      WD_LC_UNIDADES_FEDERACAO a16
    on      (a15.NR_UNIDADE_FEDERACAO                                    =
a16.NR_UNIDADE_FEDERACAO)
  join      WD_DT_DIAS    a17
    on      (a11.NR_CNPJ_ESTA_ABER_DIA_AT = a17.DT_DIA)
  join      WD_DT_ANOS    a18
    on      (a12.ANO = a18.ANO)
  join      WD_DT_ANOS    a19
    on      (a14.ANO = a19.ANO)
  join      WD_SC_SITUACOES_CADASTRAIS          a110
    on      (a11.NR_CNPJ_ESTA_SCAD_AT = a110.NR_SIT_CADASTRAL)
  join      WD_SC_MOTIVO_SIT_CADASTRAIS         a111
    on      (a11.NR_CNPJ_ESTA_SCAD_MOT_AT                                =
a111.NR_MOT_SIT_CADASTRAL)
  join      WD_DT_ANOS    a112
    on      (a17.ANO = a112.ANO)
where     (a18.NB_ANO >=  '2005'
 and a13.NR_IND_EMP_SIMEI in (2)
 and a11.NR_CNPJ_EMPA_IND_SIMP_AT in (2))
group by  a16.SG_UNIDADE_FEDERACAO,
  a17.ANO,
  a112.NB_ANO,
  a14.ANO,
  a19.NB_ANO,
  a11.NR_CNPJ_ESTA_SCAD_AT,
```

```
  a110.NM_SIT_CADASTRAL,
  a11.NR_CNPJ_ESTA_SCAD_MOT_AT,
  a111.NM_MOT_SIT_CADASTRAL

[Etapas do cálculo do mecanismo analítico:
  1. Executar crosstabbing

Número total de linhas:14741
Número total de colunas:1
Nome do servidor:10.50.243.191
ID da mensagem:4B9C963A11EAF524A6D20080EF85ADD0
ID do Trabalho640168
ID do relatório:DF17B9C011EAF395886D0080EFE56FD2
Status:Pronto
Mensagem criada:12/09/20 15:18:02
Última atualização da mensagem:12/09/20 15:18:53
```

O DataSet possui a seguinte estrutura:

Coluna	Descrição	Tipo
UF	Unidade da Federação	String
Ano_Abertura	Ano de abertura do estabelecimento	Integer
Ano_Baixa	Ano de baixa do estabelecimento. Se o estabelecimento estiver ativo o ano será ‘9999’.	Integer
Status_Cadastro	Situação cadastral do estabelecimento	String
Motivo_Baixa	Motivo da baixa do estabelecimento	String
Qtd	Quantidade de estabelecimentos	Integer
Tempo_funcionamento	Tempo de funcionamento do estabelecimento desde sua abertura até o ano de 2020	Integer

O DataSet denominado “SN”, conterá os dados das empresas optantes pelo Simples Nacional. O relatório da extração dos dados do Data Warehouse está descrito abaixo.

```
Nome do relatório:Consulta CNPJs Simples Nacional
Proprietário:ITANIMABARONI
Caminho do relatório:
CNPJ>Meus relatórios>Consulta CNPJs

Modificado:12/09/2014:58:53
```

Hora inicial:12/09/2015:22:35
Tempo de término:12/09/2015:22:46
Detalhes do Relatório
Descrição do Relatório:

Filtro do relatório:
({Empresa - 1 Ano Abertura Atual} (Nome) >= "2005") E({Empresa - Ind. Ativo SIMEI Atual} = Não:N) E ({Empresa - Ind. Op.Simples Nacional CNPJ Atual} = Sim)

Template:
Estab. - 3 UFAtual
Unidade da Federação onde se localiza o endereço doestabelecimento.
Estab. - 1 Ano Abertura Atual
Ano da data daAbertura Atual
Estab. - 1 Ano Baixa Cad. Atual
Ano da data da BaixaCad. atual
Estab. - Sit. Cad. Atual
Situação cadastral doestabelecimento
Estab. - Sit. Cad. Motivo Atual
Motivo da situaçãocadastral da estabelecimento.
Métrica:
Qtde EstabelecimentoAtual
Quantidade de Estabelecimentos.
Count({Estab. - CNPJAtual})
{~+}

Instrução SQL:

```
select      a16.SG_UNIDADE_FEDERACAO  SG_UNIDADE_FEDERACAO,
  a17.ANO  ANO,
  a112.NB_ANO  NB_ANO,
  a14.ANO  ANO0,
  a19.NB_ANO  NB_ANO0,
  a11.NR_CNPJ_ESTA_SCAD_AT  NR_SIT_CADASTRAL,
  a110.NM_SIT_CADASTRAL  NM_SIT_CADASTRAL,
  a11.NR_CNPJ_ESTA_SCAD_MOT_AT  NR_MOT_SIT_CADASTRAL,
  a111.NM_MOT_SIT_CADASTRAL  NM_MOT_SIT_CADASTRAL,
  count(distinct a11.B_CD_CNPJ_ESTA)  WJXBFS1
from      WD_CNPJ_ESTA  a11
  join      WD_DT_DIAS     a12
    on      (a11.NR_CNPJ_EMPA_ABER_DIA_AT = a12.DT_DIA)
  join      WD_PJ_ESTABELECIMENTOS      a13
    on      (a11.NR_CNPJ_ESTA = a13.NR_ESTABELECIMENTO)
  join      WD_DT_DIAS     a14
    on      (a11.NR_CNPJ_ESTA_BCAD_DIA_AT = a14.DT_DIA)
  join      WD_LC_MUNICIPIOS       a15
    on      (a11.NR_CNPJ_ESTA_LC_MC_AT = a15.NR_MUNICIPIO)
  join      WD_LC_UNIDADES_FEDERACAO a16
    on      (a15.NR_UNIDADE_FEDERACAO                                    =
a16.NR_UNIDADE_FEDERACAO)
  join      WD_DT_DIAS     a17
    on      (a11.NR_CNPJ_ESTA_ABER_DIA_AT = a17.DT_DIA)
  join      WD_DT_ANOS    a18
    on      (a12.ANO = a18.ANO)
  join      WD_DT_ANOS    a19
    on      (a14.ANO = a19.ANO)
  join      WD_SC_SITUACOES_CADASTRAIS           a110
    on      (a11.NR_CNPJ_ESTA_SCAD_AT = a110.NR_SIT_CADASTRAL)
  join      WD_SC_MOTIVO_SIT_CADASTRAIS          a111
```

```
    on      (a11.NR_CNPJ_ESTA_SCAD_MOT_AT                                    =
a111.NR_MOT_SIT_CADASTRAL)
  join      WD_DT_ANOS    a112
    on      (a17.ANO = a112.ANO)
where       (a18.NB_ANO >=  '2005'
 and a13.NR_IND_EMP_SIMEI in (2)
 and a11.NR_CNPJ_EMPA_IND_SIMP_AT in (1))
group by  a16.SG_UNIDADE_FEDERACAO,
  a17.ANO,
  a112.NB_ANO,
  a14.ANO,
  a19.NB_ANO,
  a11.NR_CNPJ_ESTA_SCAD_AT,
  a110.NM_SIT_CADASTRAL,
  a11.NR_CNPJ_ESTA_SCAD_MOT_AT,
  a111.NM_MOT_SIT_CADASTRAL

[Etapas do cálculo do mecanismo analítico:
  1. Executar crosstabbing

Número total de linhas:5533
Número total de colunas:1
Nome do servidor:10.50.243.191
ID da mensagem:EDFCF6D711EAF524E2AC0080EF750F48
ID do Trabalho640169
ID do relatório:DF17B9C011EAF395886D0080EFE56FD2
Status:Pronto
Mensagem criada:12/09/2015:22:35
Última atualização da mensagem:12/09/20 15:22:54
```

O DataSet possui a seguinte estrutura:

Coluna	Descrição	Tipo
UF	Unidade da Federação	String
Ano_Abertura	Ano de abertura do estabelecimento	Integer
Ano_Baixa	Ano de baixa do estabelecimento. Se o estabelecimento estiver ativo o ano será ‘9999’.	Integer
Status_Cadastro	Situação cadastral do estabelecimento	String
Motivo_Baixa	Motivo da baixa do estabelecimento	String
Qtd	Quantidade de estabelecimentos	Integer
Tempo_funcionamento	Tempo de funcionamento do estabelecimento desde sua abertura até o ano de 2020	Integer

O DataSet denominado “MEI”, conterá os dados das empresas optantes do MEI. O relatório da extração dos dados do Data Warehouse está descrito abaixo.

```
Nome dorelatório:Consulta CNPJs doMEI
Proprietário:ITANIMABARONI
Caminho do relatório:
CNPJ>Meus relatórios>Consulta CNPJs



Modificado:12/09/2014:58:53
Hora inicial:12/09/2015:26:31
Tempo de término:12/09/2015:26:39
Detalhes do Relatório
Descrição do Relatório:

Filtro dorelatório:
({Empresa - 1 Ano Abertura Atual} (Nome) >= "2005") E({Empresa - Ind. Ativo SIMEI Atual} = Sim:S) E ({Empresa - Ind. Op.Simples Nacional CNPJ Atual} = Não, Sim)

Template:
Estab. - 3 UFAtual
Unidade da Federação onde se localiza o endereço doestabelecimento.
Estab. - 1 Ano Abertura Atual
Ano da data daAbertura Atual
Estab. - 1 Ano Baixa Cad. Atual
Ano da data da BaixaCad. atual
Estab. - Sit. Cad. Atual
Situação cadastral doestabelecimento
Estab. - Sit. Cad. Motivo Atual
Motivo da situaçãocadastral da estabelecimento.
Métrica:
Qtde EstabelecimentoAtual
Quantidade de Estabelecimentos.
Count({Estab. - CNPJAtual})
{~+}

Instrução SQL:

select      a16.SG_UNIDADE_FEDERACAO  SG_UNIDADE_FEDERACAO,
  a17.ANO  ANO,
  a112.NB_ANO  NB_ANO,
  a14.ANO  ANO0,
  a19.NB_ANO  NB_ANO0,
  a11.NR_CNPJ_ESTA_SCAD_AT  NR_SIT_CADASTRAL,
  a110.NM_SIT_CADASTRAL  NM_SIT_CADASTRAL,
  a11.NR_CNPJ_ESTA_SCAD_MOT_AT  NR_MOT_SIT_CADASTRAL,
  a111.NM_MOT_SIT_CADASTRAL  NM_MOT_SIT_CADASTRAL,
  count(distinct a11.B_CD_CNPJ_ESTA)  WJXBFS1
from      WD_CNPJ_ESTA a11
  join      WD_DT_DIAS     a12
    on      (a11.NR_CNPJ_EMPA_ABER_DIA_AT = a12.DT_DIA)
  join      WD_PJ_ESTABELECIMENTOS      a13
    on      (a11.NR_CNPJ_ESTA = a13.NR_ESTABELECIMENTO)
  join      WD_DT_DIAS     a14
    on      (a11.NR_CNPJ_ESTA_BCAD_DIA_AT = a14.DT_DIA)
```

```
  join      WD_LC_MUNICIPIOS      a15
    on      (a11.NR_CNPJ_ESTA_LC_MC_AT = a15.NR_MUNICIPIO)
  join      WD_LC_UNIDADES_FEDERACAO a16
    on      (a15.NR_UNIDADE_FEDERACAO                                    =
a16.NR_UNIDADE_FEDERACAO)
  join      WD_DT_DIAS     a17
    on      (a11.NR_CNPJ_ESTA_ABER_DIA_AT = a17.DT_DIA)
  join      WD_DT_ANOS     a18
    on      (a12.ANO = a18.ANO)
  join      WD_DT_ANOS     a19
    on      (a14.ANO = a19.ANO)
  join      WD_SC_SITUACOES_CADASTRAIS          a110
    on      (a11.NR_CNPJ_ESTA_SCAD_AT = a110.NR_SIT_CADASTRAL)
  join      WD_SC_MOTIVO_SIT_CADASTRAIS         a111
    on      (a11.NR_CNPJ_ESTA_SCAD_MOT_AT                                =
a111.NR_MOT_SIT_CADASTRAL)
  join      WD_DT_ANOS     a112
    on      (a17.ANO = a112.ANO)
where     (a18.NB_ANO >=  '2005'
 and a13.NR_IND_EMP_SIMEI in (1)
 and a11.NR_CNPJ_EMPA_IND_SIMP_AT in (2, 1))
group by  a16.SG_UNIDADE_FEDERACAO,
  a17.ANO,
  a112.NB_ANO,
  a14.ANO,
  a19.NB_ANO,
  a11.NR_CNPJ_ESTA_SCAD_AT,
  a110.NM_SIT_CADASTRAL,
  a11.NR_CNPJ_ESTA_SCAD_MOT_AT,
  a111.NM_MOT_SIT_CADASTRAL

[Etapas do cálculo do mecanismo analítico:
  1. Executar crosstabbing

Número total de linhas:1783
Número total de colunas:1
Nome do servidor:10.50.243.191
ID da mensagem:7AE6BC1311EAF52536A40080EF15504A
ID do Trabalho640182
ID do relatório:DF17B9C011EAF395886D0080EFE56FD2
Status:Pronto
Mensagem criada:12/09/2015:26:31
Última atualização da mensagem:12/09/20 15:26:46
```

O DataSet possui a seguinte estrutura:

Coluna	Descrição	Tipo
UF	Unidade da Federação	String
Ano_Abertura	Ano de abertura do estabelecimento	Integer
Ano_Baixa	Ano de baixa do estabelecimento. Se o estabelecimento estiver ativo o ano será ‘9999’.	Integer

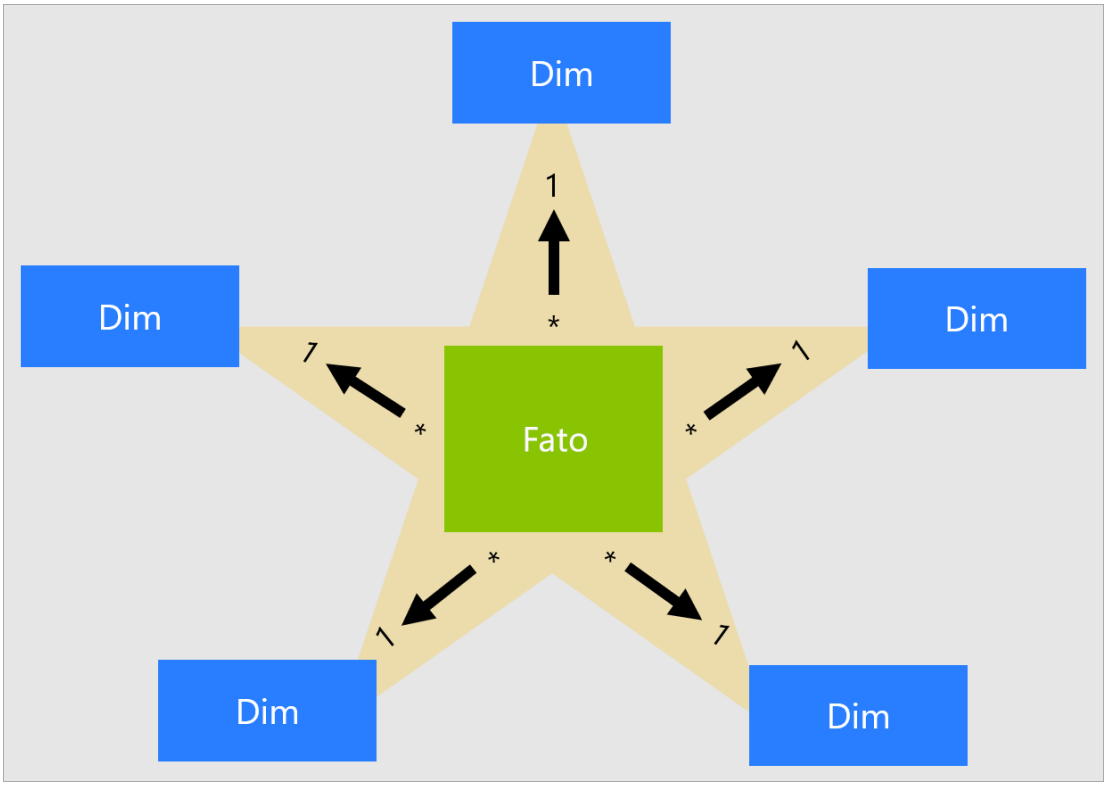
Status_Cadastro	Situação cadastral do estabelecimento	String
Motivo_Baixa	Motivo da baixa do estabelecimento	String
Qtd	Quantidade de estabelecimentos	Integer
Tempo_funcionamento	Tempo de funcionamento do estabelecimento desde sua abertura até o ano de 2020	Integer

3. Processamento/Tratamento de Dados

Os Dataset's foram extraídos de uma base dados que contem os registros de todos os CNPJ's das empresas registradas no Brasil, nacionalmente. Esta base de dados possui uma estrutura de atributos e métricas extremamente completa, incluindo ainda dados históricos.

Cabe observar que o Data Warehouse utiliza a modelagem denominada "esquema em estrela". Portanto as tabelas do modelo são classificadas como dimensão ou fato.

Resumidamente, as tabelas de dimensões descrevem as entidades de negócios, e as tabelas de fato armazenam observações ou eventos.



Nesta estrutura são considerados atributos os campos qualitativos, tais como: o número do CNPJ, Quadro societário, endereço, etc, e são consideradas métricas os campos quantitativos, tais como: quantidade de estabelecimentos, capital integralizado, etc.

Então ao extrairmos os três Dataset's, uma parte do tratamento dos dados foi feito diretamente no momento da montagem do SQL para sua obtenção. Dentre as definições para extração dos Dataset's citamos as seguintes: Foram considerados estabelecimentos com ano de abertura maior ou igual a 2005, assim teremos estabelecimentos com idade de aproximadamente 15 anos; Foram considerados todos os estados brasileiros; Incluímos a situação cadastral; Incluímos o motivo da situação cadastral, pois esta informação é importante para sabermos o motivo pelo qual o estabelecimento foi encerrado, ou pelo qual o mesmo está inativo.

Uma definição importante foi a de escolher como elemento de pesquisa o estabelecimento, e não a empresa. Isto se deve ao fato de inúmeras vezes uma empresa possui diversos estabelecimentos, e encerra a atividade em um de seus estabelecimentos, mas continua em operação, por intermédio dos outros estabelecimentos. Então, preferimos nos concentrar nas atividades dos estabelecimentos, ao invés de focarmos na empresa.

Cabe ressaltar que apesar da estrutura dos Dataset's serem a mesma, eles retratam realidades completamente diversas. No Dataset denominado "LR" teremos empresas do Lucro Presumido (receita total acima de R$ 3,6 milhões e inferior ou igual a R$ 48 milhões e que estejam liberadas da tributação por Lucro Real) e do Lucro Real (empresas com faturamento acima de R$ 78 milhões ou que exercem atividades econômicas específicas, não contempladas pelo Lucro Presumido).

As empresas do Simples Nacional, representadas pelo Dataset's nominado "SN" são Microempresas (ME) e Empresas de Pequeno Porte (EPP) com Receita Bruta anual até R$ 4,8 milhões, e representam desde empresas familiares, até empresas com um número relativamente grande de funcionários.

Quanto às empresas representadas no último Dataset, nominado de "MEI", encontram-se empresas voltadas para os profissionais autônomos que decidiram formalizar suas atividades, mas cujo faturamento anual limita-se a R$ 72.000,00.

Após a extração dos Dataset's do Data Wahehouse em arquivos "csv", foi a fase de transformá-los em um uma base de dados passíveis de serem utilizados no R. O script está incluído abaixo.

```
# library
library(readr)
library(dplyr)
library(ggplot2)
library(reshape2)
library(gridExtra)

# Diretorio de trabalho
setwd("F:/R/R-TCC")

###                Script de leitura e organização dos dados          ###
################################################################################
### EMPRESAS OPTANTES DO MEI                                          ###
################################################################################

MEI <- read_delim("F:/R/R-TCC/MEI.csv", ",", escape_double = FALSE,
            col_types = cols(ANO_ABERTURA = col_integer(),
                             ANO_BAIXA = col_integer(), QTD = col_integer()),
            trim_ws = TRUE)
MEI <- as.data.frame(MEI)
summary(MEI)
MEI$MOTIVO[MEI$MOTIVO == "NÃO INFORMADO"] <- "ATIVA"
MEI$ANO_BAIXA[is.na(MEI$ANO_BAIXA)] <- 9999
MEI <- MEI[MEI$ANO_ABERTURA <= 2019,]
summary(MEI)

################################################################################
### EMPRESAS OPTANTES PELO LUCRO REAL/PRESUMIDO                        ###
```

```
################################################################################


LR <- read_delim("F:/R/R-TCC/LR.csv", ",", escape_double = FALSE,
                 col_types = cols(ANO_ABERTURA = col_integer(),
                                  ANO_BAIXA = col_integer(), QTD = col_integer()),
                 trim_ws = TRUE)
LR <- as.data.frame(LR)
summary(LR)
LR$MOTIVO[LR$MOTIVO == "NÃO INFORMADO"] <- "ATIVA"
LR$ANO_BAIXA[is.na(LR$ANO_BAIXA)] <- 9999
LR <- LR[LR$ANO_ABERTURA <= 2019,]
summary(LR)

################################################################################
### EMPRESAS OPTANTES PELO SIMPLES NACIONAL                                  ###
################################################################################

SN <- read_delim("F:/R/R-TCC/SN.csv", ",", escape_double = FALSE,
                 col_types = cols(ANO_ABERTURA = col_integer(),
                                  ANO_BAIXA = col_integer(), QTD = col_integer()),
                 trim_ws = TRUE)
summary(SN)
SN$MOTIVO[SN$MOTIVO == "NÃO INFORMADO"] <- "ATIVA"
SN$ANO_BAIXA[is.na(SN$ANO_BAIXA)] <- 9999
SN <- SN[SN$ANO_ABERTURA <= 2019,]
summary(SN)

################################################################################

# Agrupar por UF
UF_QTD <-  MEI %>%
  group_by(UF) %>%
  summarise(QTD = sum(QTD))

barplot(UF_QTD$QTD, main = "Quantitativo de Empresas abertas do MEI",sub = "Período: 2005
- 2019", xlab = "Estados", ylab = "Quantidade",
        names.arg = UF_QTD$UF, cex.names = 0.7, col = "orange")

################################################################################

# Agrupar por UF
SN_QTD <-  SN %>%
  group_by(UF) %>%
  summarise(QTD = sum(QTD))

barplot(SN_QTD$QTD, main = "Quantitativo de Empresas abertas do Simples Nacional",sub =
"Período: 2005 - 2019", xlab = "Estados", ylab = "Quantidade",
        names.arg = SN_QTD$UF, cex.names = 0.7, col = "blue")

################################################################################
# Agrupar por UF
LR_QTD <-  LR %>%
  group_by(UF) %>%
  summarise(QTD = sum(QTD))
```

```
barplot(LR_QTD$QTD, main = "Quantitativo de Empresas abertas do Lucro Real/Presumido",sub
= "Período: 2005 - 2019", xlab = "Estados", ylab = "Quantidade",
        names.arg = LR_QTD$UF, cex.names = 0.7, col = "RED")


################################################################################

MEI$ANO_BAIXA[MEI$ANO_BAIXA == 9999] <- 2020
MEI$TEMPO_FUNCIONAMENTO <- MEI$ANO_BAIXA - MEI$ANO_ABERTURA
MEI$ANO_BAIXA[MEI$SIT_CADASTRAL == "Ativa"] <- 9999

################################################################################

SN$ANO_BAIXA[SN$ANO_BAIXA == 9999] <- 2020
SN$TEMPO_FUNCIONAMENTO <- SN$ANO_BAIXA - SN$ANO_ABERTURA
SN$ANO_BAIXA[SN$SIT_CADASTRAL == "Ativa"] <- 9999

################################################################################

LR$ANO_BAIXA[LR$ANO_BAIXA == 9999] <- 2020
LR$TEMPO_FUNCIONAMENTO <- LR$ANO_BAIXA - LR$ANO_ABERTURA
LR$ANO_BAIXA[LR$SIT_CADASTRAL == "Ativa"] <- 9999
LR <- LR[LR$TEMPO_FUNCIONAMENTO >= 0,]
################################################################################
################################################################################
################################################################################

MEI_2020$ANO_BAIXA[MEI_2020$ANO_BAIXA == 9999] <- 2020
MEI_2020$TEMPO_FUNCIONAMENTO <- MEI_2020$ANO_BAIXA - MEI_2020$ANO_ABERTURA
MEI_2020$ANO_BAIXA[MEI_2020$SIT_CADASTRAL == "Ativa"] <- 9999

################################################################################

SN_2020$ANO_BAIXA[SN_2020$ANO_BAIXA == 9999] <- 2020
SN_2020$TEMPO_FUNCIONAMENTO <- SN_2020$ANO_BAIXA - SN_2020$ANO_ABERTURA
SN_2020$ANO_BAIXA[SN_2020$SIT_CADASTRAL == "Ativa"] <- 9999

################################################################################

LR_2020$ANO_BAIXA[LR_2020$ANO_BAIXA == 9999] <- 2020
LR_2020$TEMPO_FUNCIONAMENTO <- LR_2020$ANO_BAIXA - LR_2020$ANO_ABERTURA
LR_2020$ANO_BAIXA[LR_2020$SIT_CADASTRAL == "Ativa"] <- 9999

################################################################################
```

Após a execução do script de leitura dos dados, verificamos que alguns campos necessitariam de ajustes.

O primeiro campo foi o ano da baixa (ano_baixa) que no ato de extração do DW este veio preenchido com o valor “Não se aplica” quando o estabelecimento ainda está ativo. Então optamos por alterar este valor para “9999”, simulando uma data de baixa infinita, já que o estabelecimento está ativo.

O outro campo que necessitou de ajuste foi o motivo da baixa (motivo), pois este veio preenchido com “NÃO INFORMADO” quando o estabelecimento está ativo. Então alteramos este valor para “Ativo”, informando que se o estabelecimento está ativo o motivo da baixa deverá ser “Ativo”.

Outras verificações foram feitas a fim de testarmos a integridade dos dados, a saber: ano de abertura menor que ano da baixa – sem ocorrências; quantidade de estabelecimentos menor ou igual a zero – sem ocorrências; situação cadastral e motivo da baixa sem preenchimento quando o ano de baixa for maior que zero – sem ocorrências.

Cabe observar que a inclusão do atributo tempo de funcionamento foi efetuada pelo script de tratamento dos dados, e não extraído diretamente da base do DW.

Concluídas as verificações de integridades efetuamos a importação dos dados para o R. Os resultados encontram-se abaixo, onde por intermédio da função “summary”, exibiremos os resultados antes, e depois das interações com os Dataset’s.

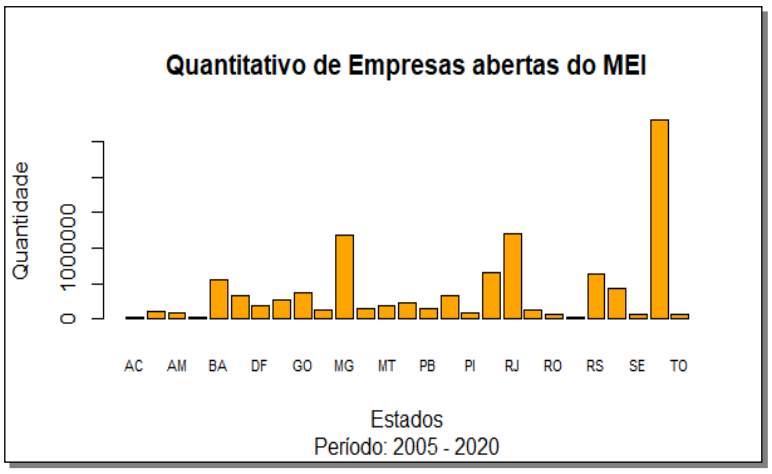
A fim de efetuarmos as importações dos dados, e suas conversões em objetos passíveis de utilização no R, assim como os gráficos exibidos abaixo foi necessária a instalação dos pacotes readr, dplyr, ggplot2, reshape2 e gridExtra. O script de importação foi nomeado como “main.R”.

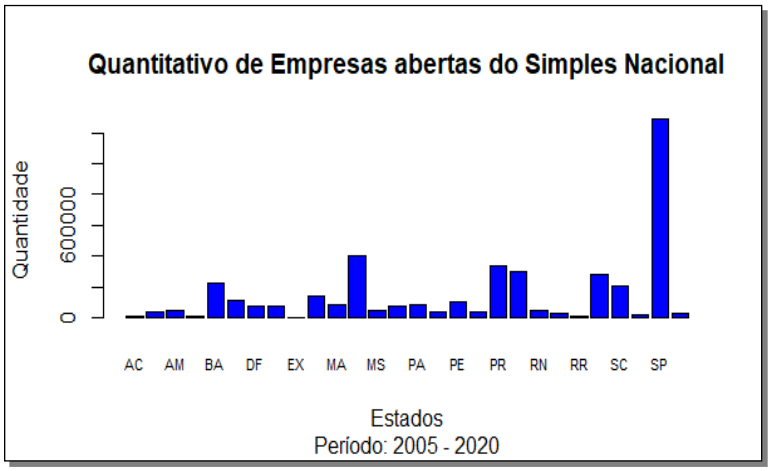
```
> ################################################################################
>
> MEI <- read_delim("F:/R/MEI.csv", ",", escape_double = FALSE,
+                   col_types = cols(ANO_ABERTURA = col_integer(),
+                                    ANO_BAIXA = col_integer(), QTD = col_integer()),
+                   trim_ws = TRUE)
Warning: 967 parsing failures.
row       col   expected         actual           file
  1 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/MEI.csv'
  2 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/MEI.csv'
  3 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/MEI.csv'
  4 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/MEI.csv'
  5 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/MEI.csv'
... ......... .......... .............. ..............
See problems(...) for more details.

> MEI <- as.data.frame(MEI)
> summary(MEI)
      UF              ANO_ABERTURA    ANO_BAIXA    SIT_CADASTRAL         MOTIVO              QTD
 Length:1783        Min.   :2005   Min.   :2006   Length:1783        Length:1783        Min.   :     1
 Class :character   1st Qu.:2009   1st Qu.:2013   Class :character   Class :character   1st Qu.:     1
 Mode  :character   Median :2011   Median :2016   Mode  :character   Mode  :character   Median :     2
                    Mean   :2012   Mean   :2015                                         Mean   :  5816
                    3rd Qu.:2015   3rd Qu.:2018                                         3rd Qu.:    20
                    Max.   :2020   Max.   :2020                                         Max.   :615618
                                   NA's   :967
> MEI$MOTIVO[MEI$MOTIVO == "NÃO INFORMADO"] <- "ATIVA"
> MEI$ANO_BAIXA[is.na(MEI$ANO_BAIXA)] <- 9999
> summary(MEI)
      UF              ANO_ABERTURA    ANO_BAIXA    SIT_CADASTRAL         MOTIVO              QTD
 Length:1783        Min.   :2005   Min.   :2006   Length:1783        Length:1783        Min.   :     1
 Class :character   1st Qu.:2009   1st Qu.:2017   Class :character   Class :character   1st Qu.:     1
 Mode  :character   Median :2011   Median :9999   Mode  :character   Mode  :character   Median :     2
                    Mean   :2012   Mean   :6345                                         Mean   :  5816
                    3rd Qu.:2015   3rd Qu.:9999                                         3rd Qu.:    20
                    Max.   :2020   Max.   :9999                                         Max.   :615618
>
> ################################################################################
```

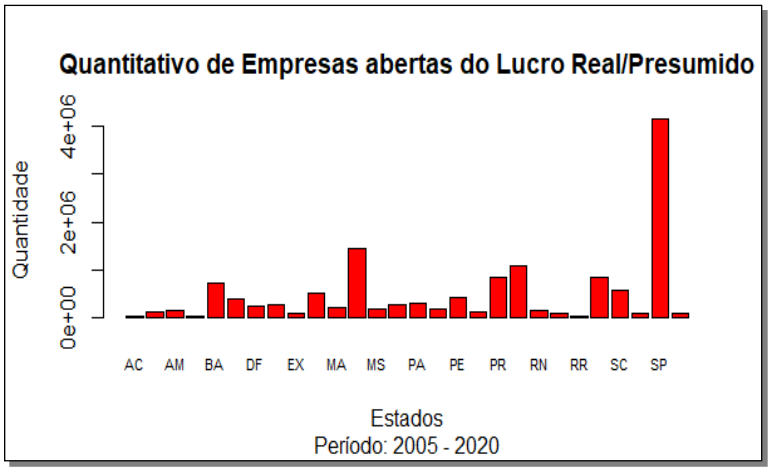
```
> ################################################################################
>
> SN <- read_delim("F:/R/SN.csv", ",", escape_double = FALSE,
+                  col_types = cols(ANO_ABERTURA = col_integer(),
+                                   ANO_BAIXA = col_integer(), QTD = col_integer()),
+                  trim_ws = TRUE)
Warning: 2049 parsing failures.
row       col   expected         actual          file
  1 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/SN.csv'
  2 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/SN.csv'
  3 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/SN.csv'
  4 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/SN.csv'
  5 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/SN.csv'
... ......... .......... .............. .............
See problems(...) for more details.

> summary(SN)
      UF              ANO_ABERTURA    ANO_BAIXA    SIT_CADASTRAL         MOTIVO              QTD
 Length:5532        Min.   :2005   Min.   :2005   Length:5532        Length:5532        Min.   :     1.0
 Class :character   1st Qu.:2008   1st Qu.:2012   Class :character   Class :character   1st Qu.:     2.0
 Mode  :character   Median :2010   Median :2015   Mode  :character   Mode  :character   Median :     6.0
                    Mean   :2011   Mean   :2015                                         Mean   :   753.8
                    3rd Qu.:2014   3rd Qu.:2018                                         3rd Qu.:    21.0
                    Max.   :2020   Max.   :2020                                         Max.   :165117.0
                                   NA's   :2049
> SN$MOTIVO[SN$MOTIVO == "NÃO INFORMADO"] <- "ATIVA"
> SN$ANO_BAIXA[is.na(SN$ANO_BAIXA)] <- 9999
> summary(SN)
      UF              ANO_ABERTURA    ANO_BAIXA    SIT_CADASTRAL         MOTIVO              QTD
 Length:5532        Min.   :2005   Min.   :2005   Length:5532        Length:5532        Min.   :     1.0
 Class :character   1st Qu.:2008   1st Qu.:2014   Class :character   Class :character   1st Qu.:     2.0
 Mode  :character   Median :2010   Median :2019   Mode  :character   Mode  :character   Median :     6.0
                    Mean   :2011   Mean   :4972                                         Mean   :   753.8
                    3rd Qu.:2014   3rd Qu.:9999                                         3rd Qu.:    21.0
                    Max.   :2020   Max.   :9999                                         Max.   :165117.0
>
> ################################################################################
```

```
> ##############################################################################
>
> LR <- read_delim("F:/R/NORMAL.csv", ",", escape_double = FALSE,
+                  col_types = cols(ANO_ABERTURA = col_integer(),
+                                   ANO_BAIXA = col_integer(), QTD = col_integer()),
+                  trim_ws = TRUE)
Warning: 3659 parsing failures.
row       col   expected         actual             file
  1 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/NORMAL.csv'
  2 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/NORMAL.csv'
  3 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/NORMAL.csv'
  4 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/NORMAL.csv'
  5 ANO_BAIXA an integer NÃ£o se aplica 'F:/R/NORMAL.csv'
... ......... .......... ............... ..................
See problems(...) for more details.

> LR <- as.data.frame(LR)
> summary(LR)
      UF              ANO_ABERTURA    ANO_BAIXA    SIT_CADASTRAL         MOTIVO              QTD
 Length:14718       Min.   :2005   Min.   :2005   Length:14718       Length:14718       Min.   :     1.0
 Class :character   1st Qu.:2007   1st Qu.:2012   Class :character   Class :character   1st Qu.:     2.0
 Mode  :character   Median :2010   Median :2015   Mode  :character   Mode  :character   Median :    10.0
                    Mean   :2010   Mean   :2015                                         Mean   :   937.9
                    3rd Qu.:2013   3rd Qu.:2018                                         3rd Qu.:   207.0
                    Max.   :2020   Max.   :2020                                         Max.   :188585.0
                                   NA's   :3659
> LR$MOTIVO[LR$MOTIVO == "NÃO INFORMADO"] <- "ATIVA"
> LR$ANO_BAIXA[is.na(LR$ANO_BAIXA)] <- 9999
> summary(LR)
      UF              ANO_ABERTURA    ANO_BAIXA    SIT_CADASTRAL         MOTIVO              QTD
 Length:14718       Min.   :2005   Min.   :2005   Length:14718       Length:14718       Min.   :     1.0
 Class :character   1st Qu.:2007   1st Qu.:2013   Class :character   Class :character   1st Qu.:     2.0
 Mode  :character   Median :2010   Median :2017   Mode  :character   Mode  :character   Median :    10.0
                    Mean   :2010   Mean   :4000                                         Mean   :   937.9
                    3rd Qu.:2013   3rd Qu.:2020                                         3rd Qu.:   207.0
                    Max.   :2020   Max.   :9999                                         Max.   :188585.0
>
> ##############################################################################
```

Os gráficos abaixo têm por objetivo demonstrar o quantitativo de estabelecimentos abertos por estado, no período de estudo – anos 2005 a 2020.







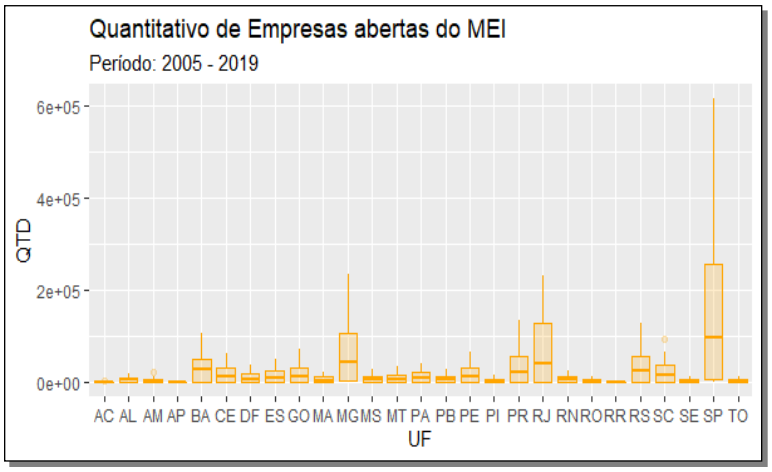
Antes de prosseguirmos para a análise dos dados, que será efetuada no próximo tópico, cabe esclarecer o que seria a unidade da federação denominada “EX”.

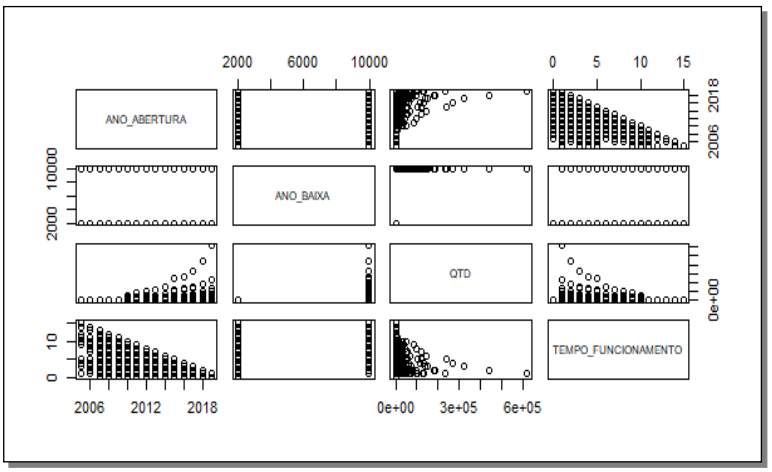
Algumas empresas tem sua sede no exterior, e efetuam a abertura de empresas no Brasil a fim de desempenharem determinadas operações comerciais. Um exemplo são as Tradings.

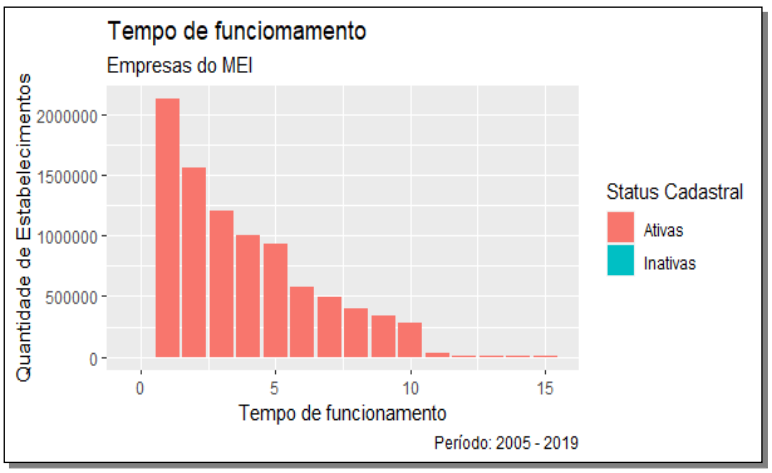
Segundo ponto que pode ser observado pela simples análise dos gráficos acima é a de que o estado de São Paulo, de certo modo é um Outlier, pois o número de estabelecimentos abertos, em todas as categorias, supera sobremaneira os outros estados da federação.

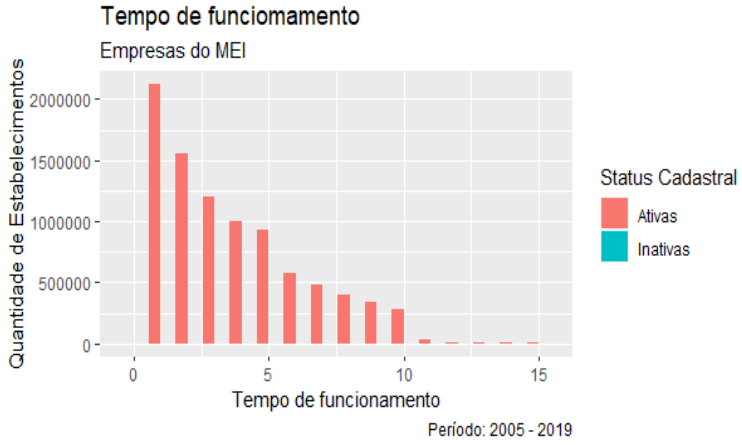
4. Análise e Exploração dos Dados

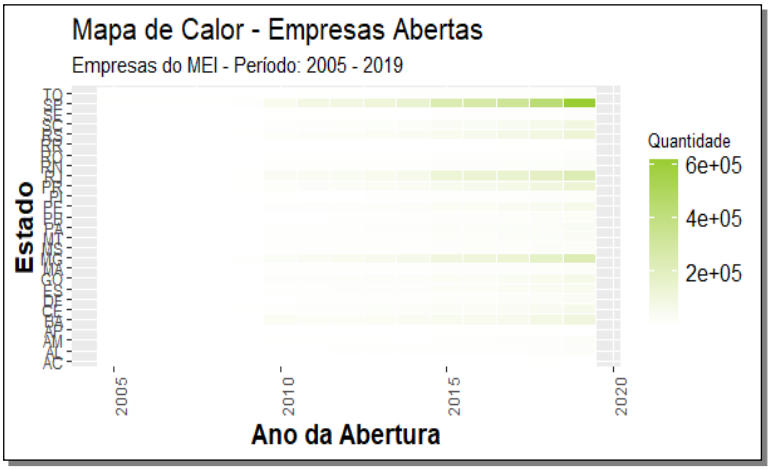
A análise dos dados foi iniciada verificando-se a distribuição do quantitativo de estabelecimentos abertos, por estado no período aqui proposto: 2005 a 2019.

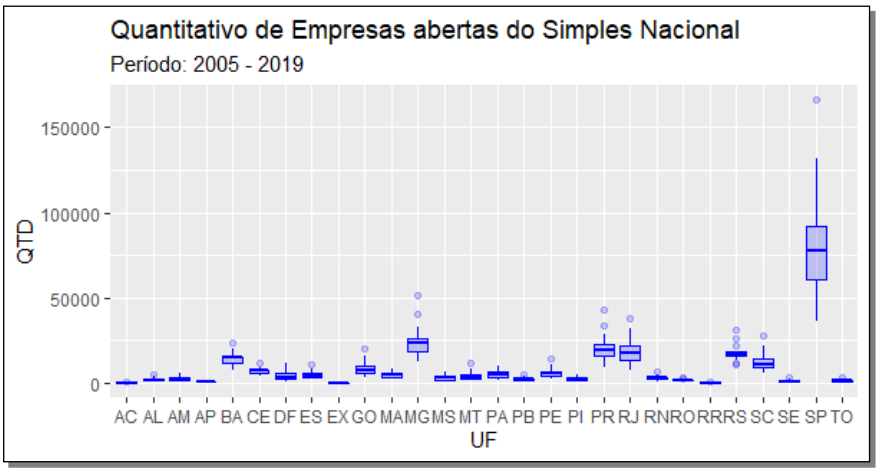


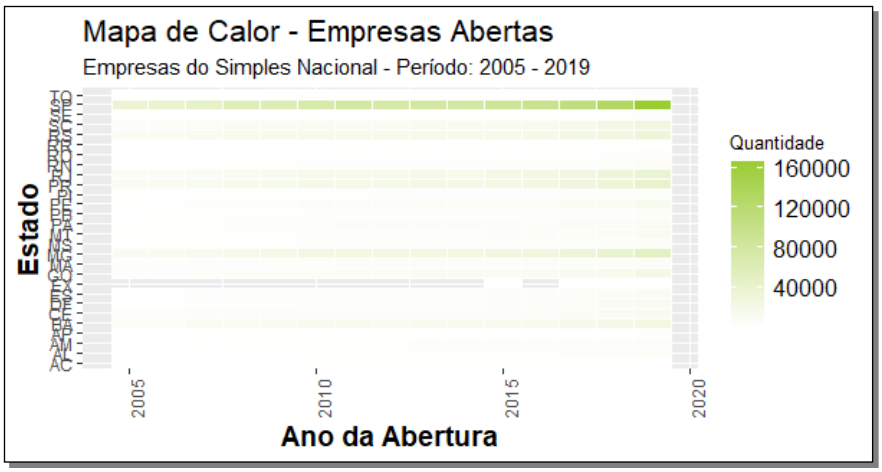


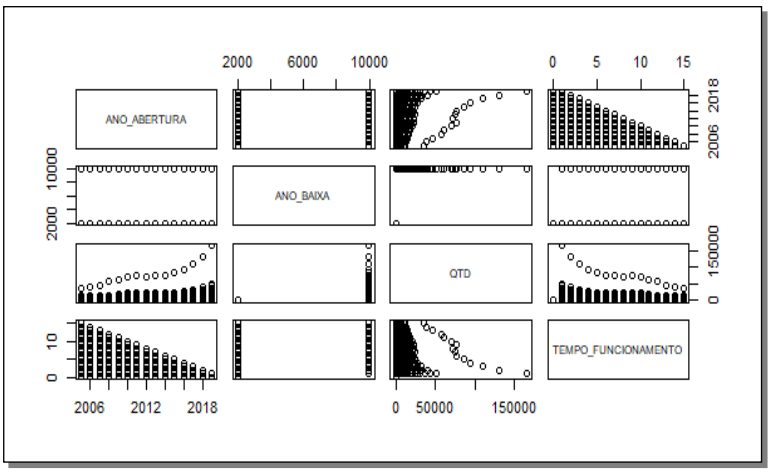


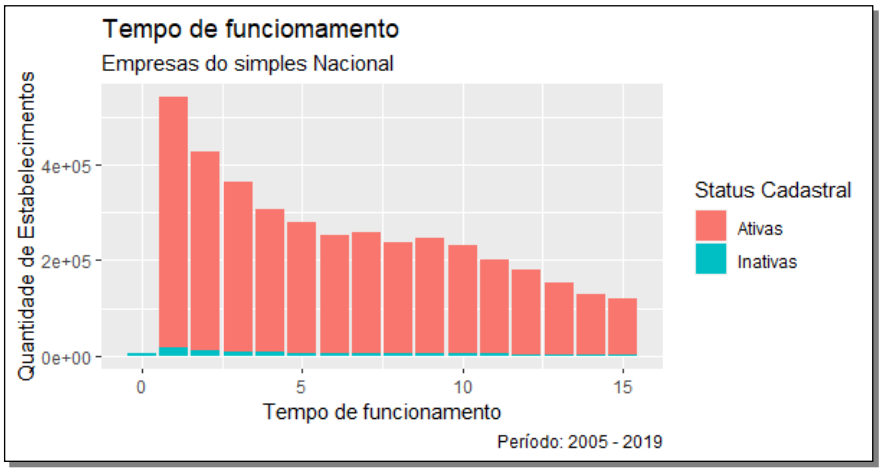

Tempo de funciomamento
Empresas do MEI
Quantidade de Estabelecimentos
2000000
1500000
1000000
500000
0
0
5
10
15
Tempo de funcionamento
Período: 2005 - 2019
Status Cadastral
Ativas
Inativas

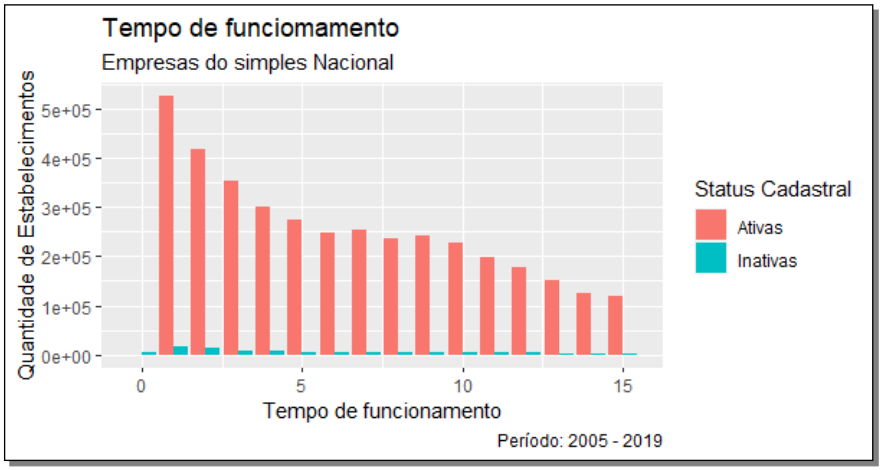

Mapa de Calor - Empresas Abertas
Empresas do MEI - Período: 2005 - 2019
Estado
2005
2010
2015
2020
Ano da Abertura
Quantidade
6e+05
4e+05
2e+05

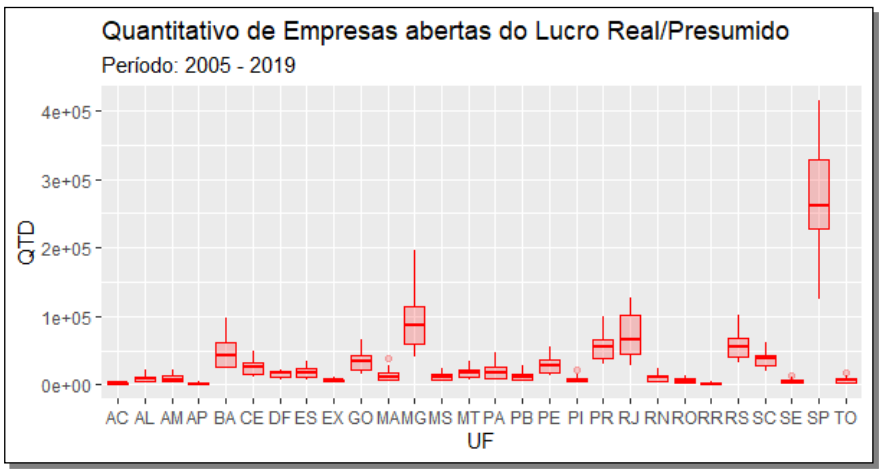

Quantitativo de Empresas abertas do Simples Nacional
Período: 2005 - 2019
QTD
150000
100000
50000
0
AC AL AM AP BA CE DF ES EX GO MA MG MS MT PA PB PE PI PR RJ RN RO RR RS SC SE SP TO
UF

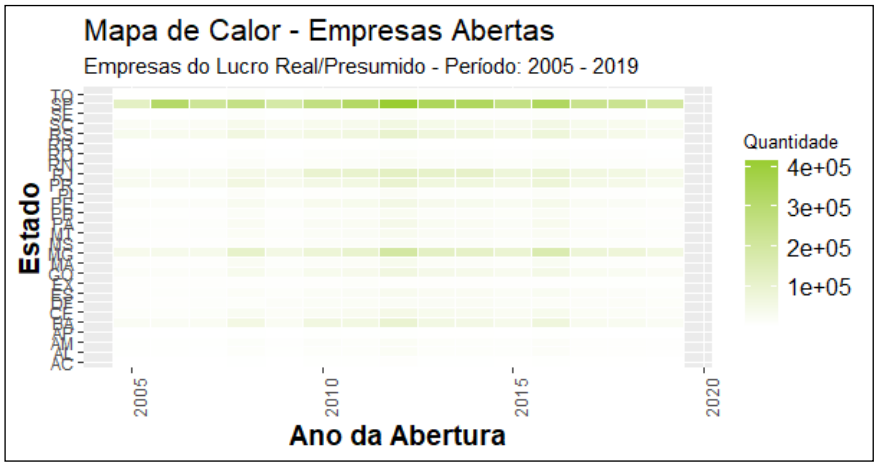

Mapa de Calor - Empresas Abertas
Empresas do Simples Nacional - Período: 2005 - 2019
Estado
Quantidade
160000
120000
80000
40000
2005
2010
2015
2020
Ano da Abertura

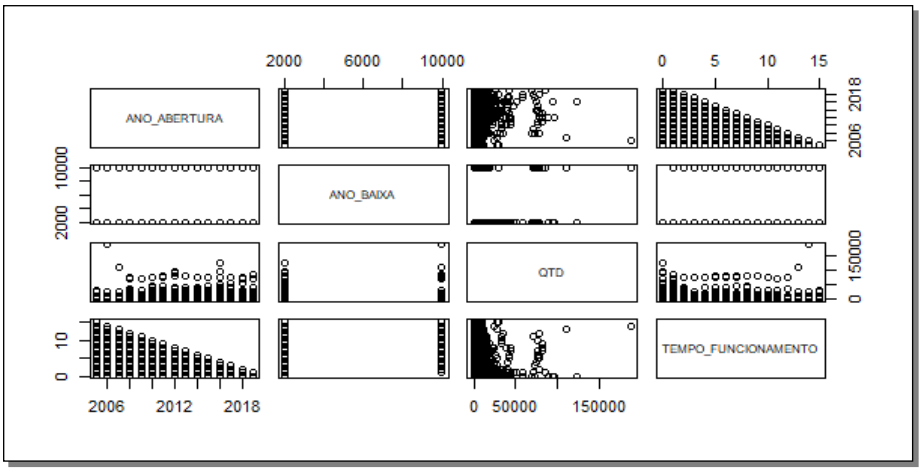

ANO_ABERTURA
ANO_BAIXA
QTD
TEMPO_FUNCIONAMENTO
2000
6000
10000
0
5
10
15
2006
2012
2018
0
50000
150000

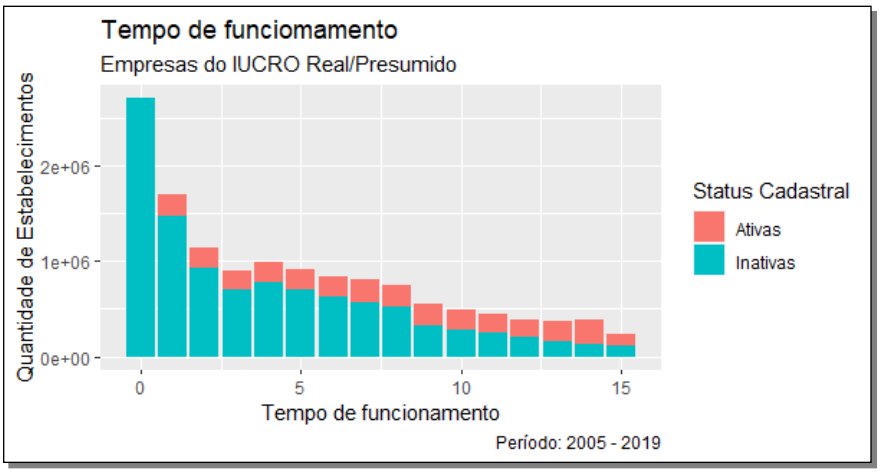

Tempo de funciomamento
Empresas do simples Nacional
Quantidade de Estabelecimentos
4e+05
2e+05
0e+00
0
5
10
15
Tempo de funcionamento
Período: 2005 - 2019
Status Cadastral
Ativas
Inativas

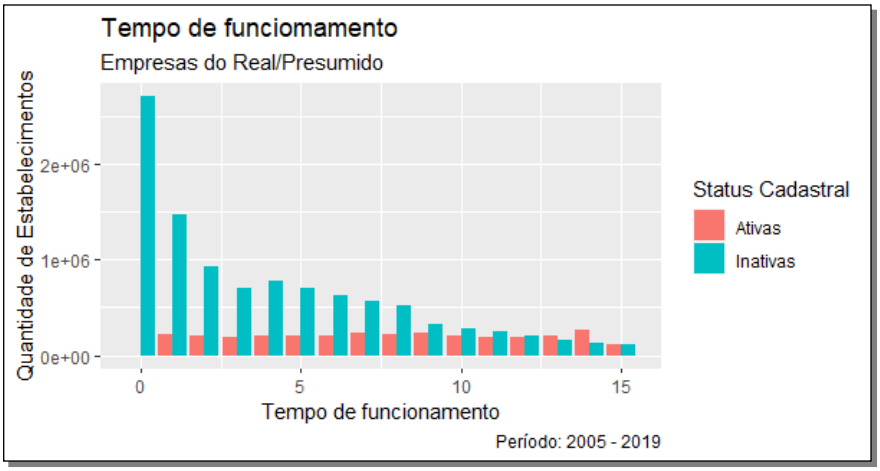

Tempo de funciomamento
Empresas do simples Nacional
Quantidade de Estabelecimentos
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
0
5
10
15
Tempo de funcionamento
Período: 2005 - 2019
Status Cadastral
Ativas
Inativas


Quantitativo de Empresas abertas do Lucro Real/Presumido
Período: 2005 - 2019
QTD
4e+05
3e+05
2e+05
1e+05
0e+00
AC AL AM AP BA CE DF ES EX GO MA MG MS MT PA PB PE PI PR RJ RN RO RR RS SC SE SP TO
UF

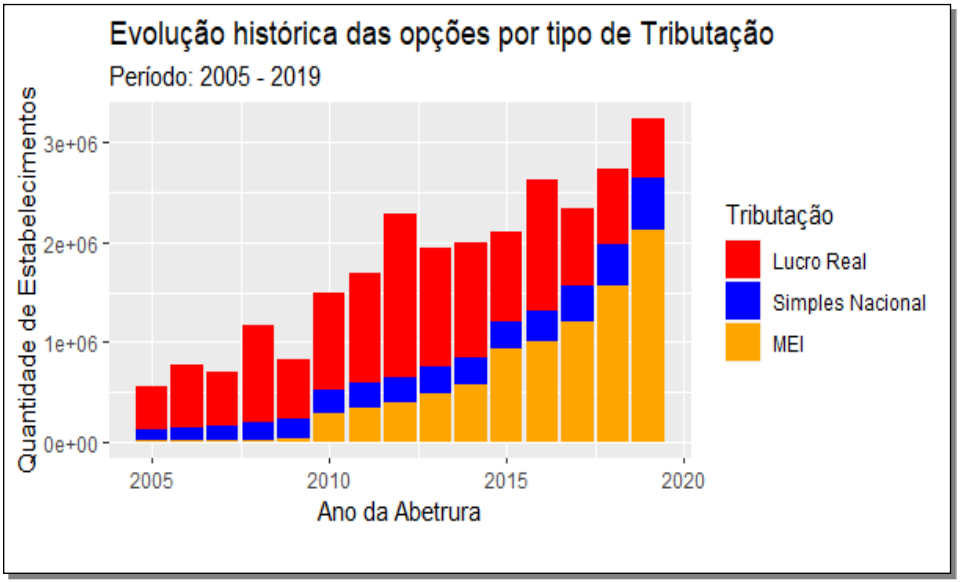

Mapa de Calor - Empresas Abertas
Empresas do Lucro Real/Presumido - Período: 2005 - 2019
Estado
Quantidade
4e+05
3e+05
2e+05
1e+05
2005
2010
2015
2020
Ano da Abertura

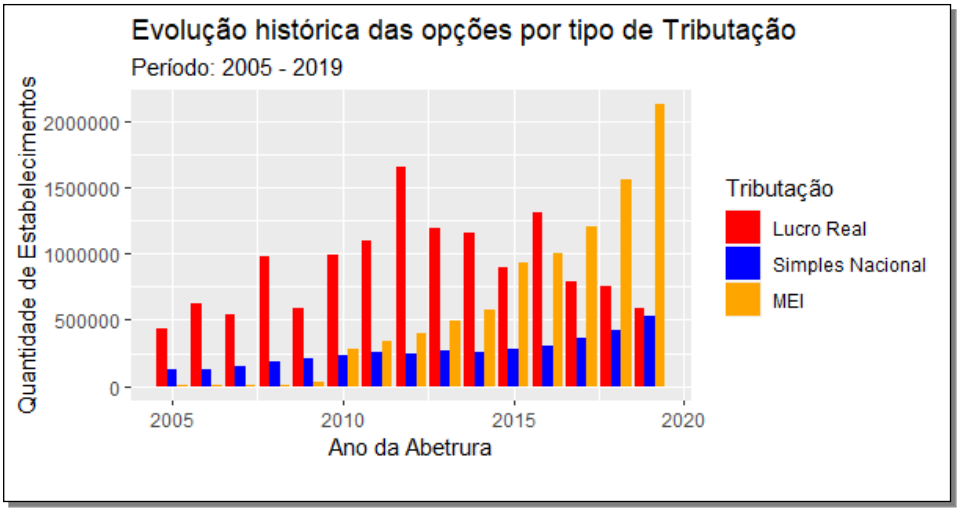

ANO_ABERTURA
ANO_BAIXA
QTD
TEMPO_FUNCIONAMENTO
2000
6000
10000
0
5
10
15
2006
2012
2018
0
50000
150000


Tempo de funciomamento
Empresas do IUCRO Real/Presumido
Quantidade de Estabelecimentos
2e+06
1e+06
0e+00
0
5
10
15
Tempo de funcionamento
Período: 2005 - 2019
Status Cadastral
Ativas
Inativas



Na segunda análise avaliamos a evolução do quantitativo de estabelecimentos abertos, no período em estudo.





A observação dos gráficos acima nos mostra de forma bastante clara, a opção pelo Regime de Tributação - MEI, em detrimento do regime pelo Lucro Real/Presumido, enquanto o regime do Simples Nacional permaneceu praticamente inalterado.

Estes gráficos, por si só, nos remetem a inúmeras discussões econômico-sociais que fogem do escopo do presente estudo, mas que não podemos deixar de apontar, principalmente por que a observância dos dados não pode se ater a números, mas a informação trazida pelos mesmos.

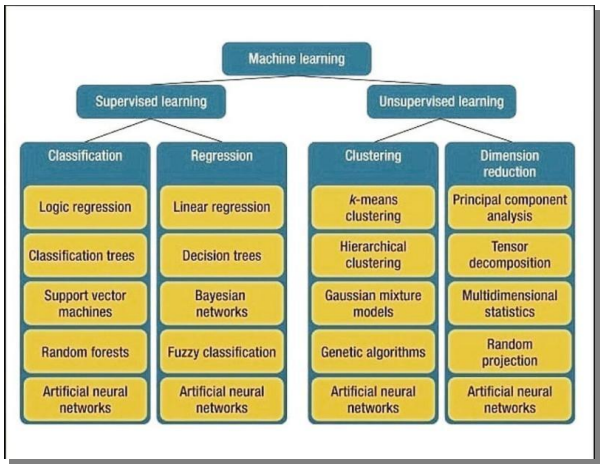
Então, algumas indagações que poderiam ser feitas, dentro do escopo da análise do tempo de sobrevivência das empresas:

- O comportamento do ciclo de vida das empresas é afetado pelo tipo de tributação, ou independe do mesmo?
- A opção pelo regime-MEI é fruto de uma mudança de cultura pelo empreendedorismo, ou é a denominada “pejotização” das relações trabalhistas?
- A retração do número de estabelecimentos pelo regime do Lucro Real/Presumido reflete a eterna crise econômica de onde o Brasil nunca saiu, haja vista, serem regimes de empresas de maior porte, com risco elevado do retorno do valor investido, vinculado à instabilidade econômica do País?

Estes, e tantos outros questionamentos devem fazer parte da análise dos dados, isto porque com já dito, os dados devem ser vistos como informação, e principalmente, no objeto do presente trabalho, os dados/informação devem ser analisados em conjunto com os aspectos externos – aspectos político/sociais/econômicos/mercadológicos etc.

5. Criação de Modelos de Machine Learning

Antes de prosseguirmos devemos lembrar que os métodos de maching learning são divididos em: Aprendizagem supervisionada, Aprendizagem não-supervisionada, Aprendizado semisupervisionado e Aprendizagem por esforço.



Para atingirmos o objetivo do trabalho utilizaremos duas técnicas de Maching Learning: Árvore de decisão e Mineração de texto. A escolha por estes métodos deve-se ao fato da facilidade de interpretação, características dos Dataset's e objetivo do problema proposto.

Para o desenvolvimento do algoritmo de treinamento da árvore de decisão utilizamos o pacote "rpart" e "rpart.plot" do R, cujas funcionalidades permitem a criação de árvores de classificação, regressão e sobrevivência. Conforme descrito no manual de referência, este pacote é uma implementação baseada no livro de "Classification and Regression Trees", de 1984, dos autores Breiman, Friedman, Olshen e Stone.

Apresentamos o script utilizado na modelagem, iniciando com o carregamento das bibliotecas necessárias (library's no R).

```
###########
library(readr)
library(dplyr)
library(ggplot2)
library(reshape2)
library(gridExtra)
############################################################
# BIBLIOTECAS PARA A GERAÇÃO
# DAS ÁRVORES DE CLASSIFICAÇÃO
############################################################
library(rpart)
library(rpart.plot)
library(xtable)
```

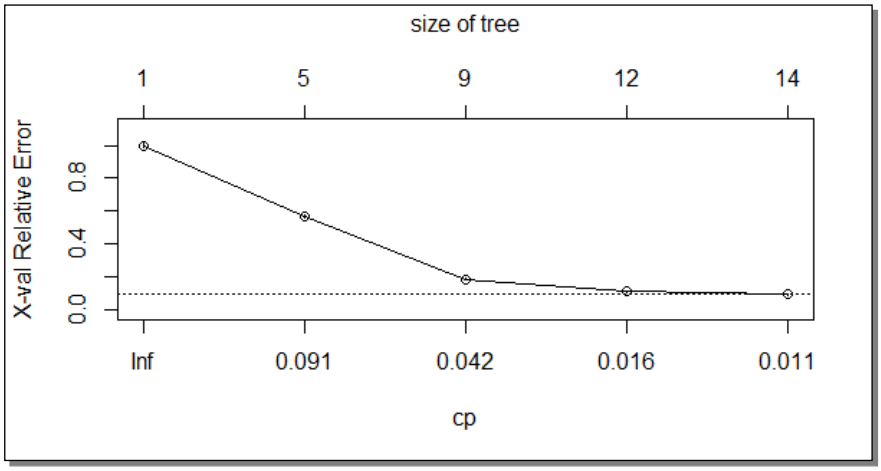
Para otimizarmos o algoritmo incluímos um variável nos DataSet's denominada Cod_motivo, que nada mais é do que um índice para os Motivos da baixa dos estabelecimentos (ver apêndice). Desta forma nossos DataSet's poderão ser trabalhados de forma mais eficiente. Lembrando que este campo é uma string contendo os motivos das situações cadastrais.

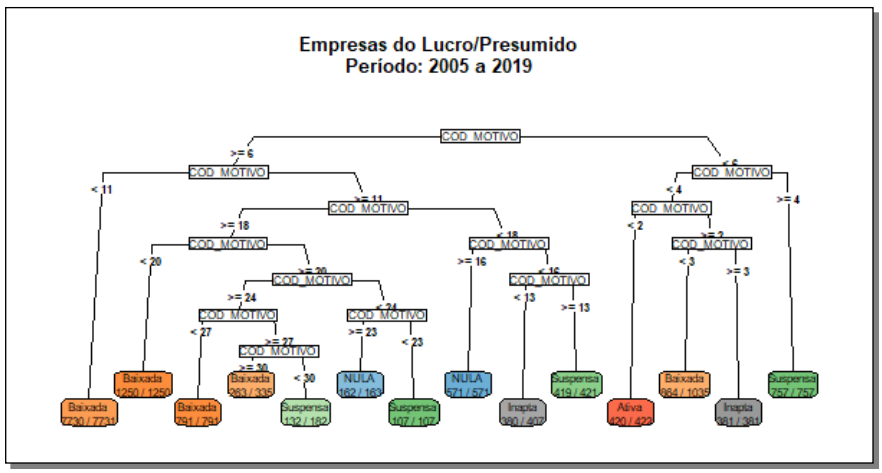
```
#############################################################################
# SCRIPT PARA A GERAÇÃO DAS ÁRVORES DE CLASSIFICAÇÃO
#############################################################################
############### PARA O LUCRO REAL/PRESUMIDO
#############################################################################
LR_MOTIVO <- as.data.frame(unique(LR$MOTIVO))
i <- row(LR_MOTIVO[1])
for (a in i){
  LR_MOTIVO$Desc[a] <- a
}

names(LR_MOTIVO) <- c("MOTIVO","COD_MOTIVO")
LR1 <- inner_join(LR,LR_MOTIVO,by="MOTIVO")
############### TESTE PARA DIFERENTES HIPÓTESES DE ÁRVORES
#############################################################################
# Modelo 1
```

```
#arvore <- rpart(SIT_CADASTRAL ~ COD_MOTIVO + QTD ,
                data = LR1, method = "class")


# Modelo 2
#arvore <- rpart(SIT_CADASTRAL ~ TEMPO_FUNCIONAMENTO + QTD ,
            data = LR1, method = "class")
# Modelo3
#arvore <- rpart(SIT_CADASTRAL ~ QTD + TEMPO_FUNCIONAMENTO ,
            data = LR1, method = "class")


################################################################################
############### PARA O LUCRO REAL/PRESUMIDO
################################################################################

printcp(arvore)         # Exibe os resultados
plotcp(arvore)          # visualizar os resultados de cross-validation
summary(arvore)         # detalhamento da montagem da árvore)

rpart.plot(prune(arvore, cp = 0.010000),
         tweak = 1.1,
         type = 5,
         branch = .9,
         yesno = F,
         extra = 2, #under = T,
         legend.x = NA,
         gap = 0,space = 0,
#          shadow.col = "GRAY",
         main = "Empresas do Lucro/Presumido\nPeríodo: 2005 a 2019\n")

rm(a,i,LR1,arvore,LR_MOTIVO)
################################################################################
```

Após esta alteração no Dataset efetuamos a primeira simulação denominada “Modelo 1”, com os resultados demonstrados abaixo.

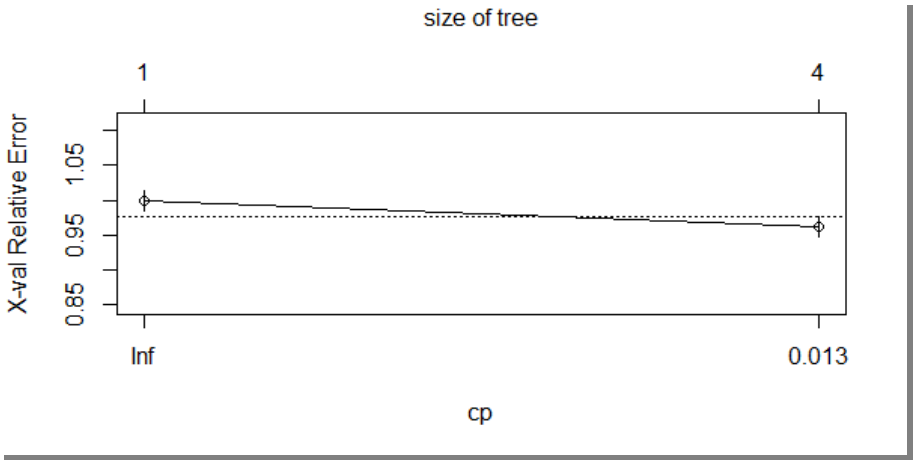
```
> printcp(arvore)            # Exibe os resultados

Classification tree:
rpart(formula = SIT_CADASTRAL ~ COD_MOTIVO + QTD, data = LR1,
    method = "class")

Variables actually used in tree construction:
[1] COD_MOTIVO

Root node error: 3602/14553 = 0.24751

n= 14553

        CP nsplit rel error   xerror      xstd
1 0.105081      0  1.000000 1.000000 0.0144537
2 0.078984      4  0.567463 0.567463 0.0116368
3 0.022349      8  0.187674 0.187674 0.0070486
4 0.011383     11  0.113270 0.113270 0.0055286
5 0.010000     13  0.090505 0.094392 0.0050590
```

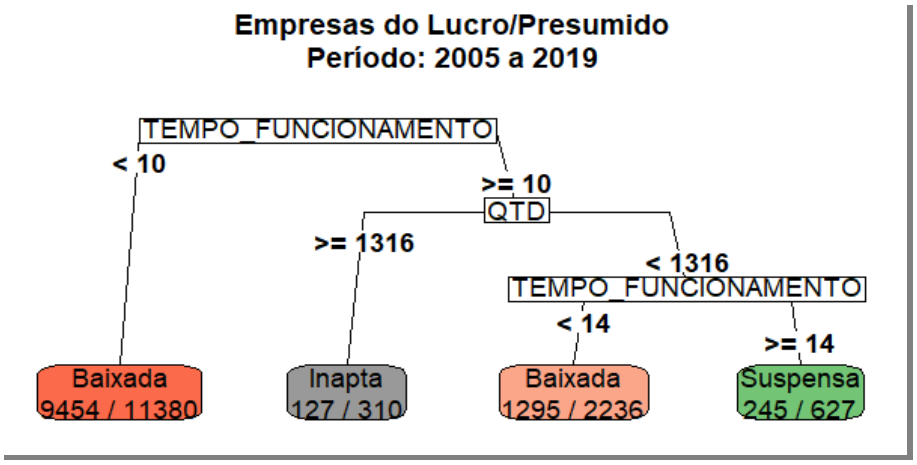




Os resultados do “Modelo 2”, são demonstrados abaixo.

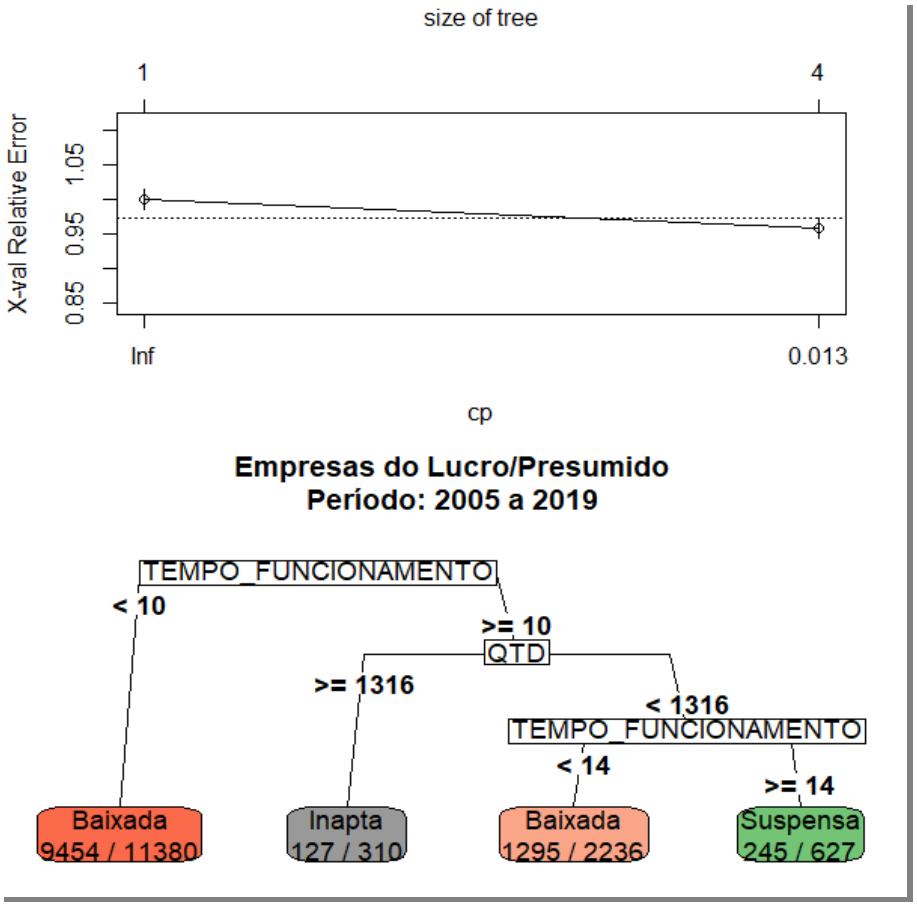
```
> printcp(arvore)          # Exibe os resultados

Classification tree:
rpart(formula = SIT_CADASTRAL ~ TEMPO_FUNCIONAMENTO + QTD, data = LR1,
    method = "class")

Variables actually used in tree construction:
[1] QTD                 TEMPO_FUNCIONAMENTO

Root node error: 3602/14553 = 0.24751

n= 14553

        CP nsplit rel error  xerror     xstd
1 0.015732      0    1.0000 1.00000 0.014454
2 0.010000      3    0.9528 0.96141 0.014262
```





Por fim, o “Modelo 3”:

```
> printcp(arvore)           # Exibe os resultados

Classification tree:
rpart(formula = SIT_CADASTRAL ~ QTD + TEMPO_FUNCIONAMENTO, data = LR1,
    method = "class")

Variables actually used in tree construction:
[1] QTD                 TEMPO_FUNCIONAMENTO

Root node error: 3602/14553 = 0.24751

n= 14553

        CP nsplit rel error  xerror     xstd
1 0.015732      0    1.0000 1.00000 0.014454
2 0.010000      3    0.9528 0.95919 0.014250
```

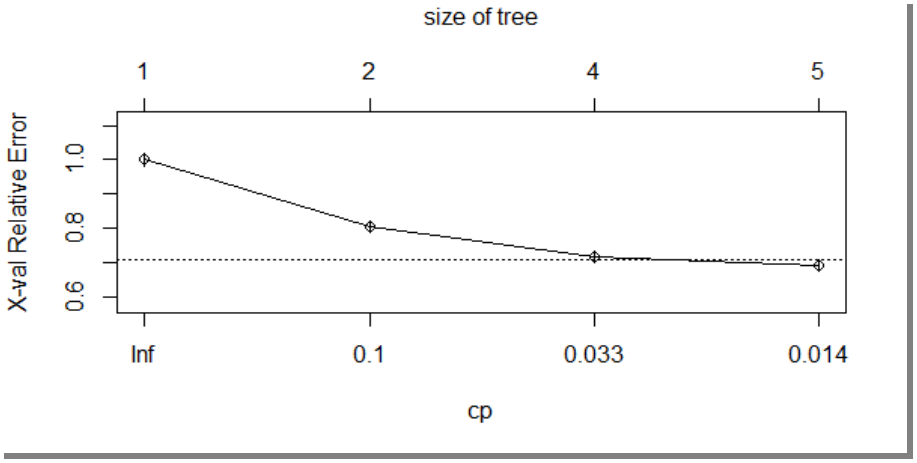


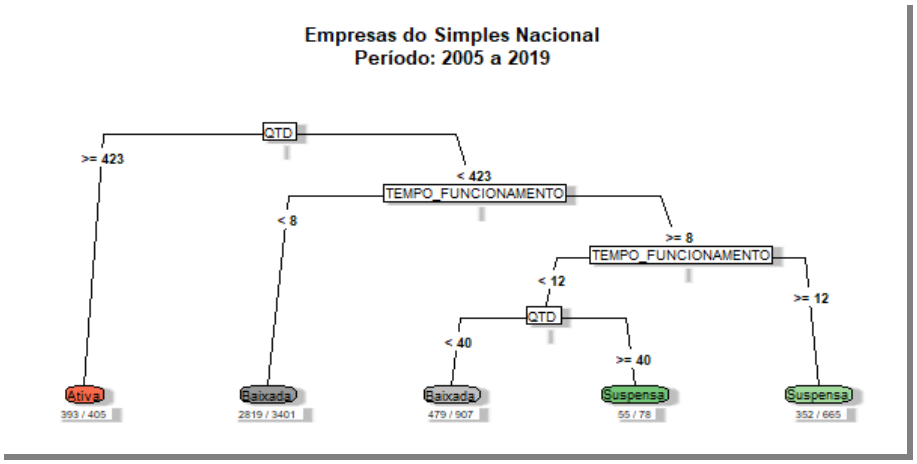
O script, assim como os resultados, para as Empresas optantes pelo Simples Nacional está demonstrado abaixo.

```
################################################################################
################ PARA O SN
```

```
################################################################################
##############

SN_MOTIVO <- as.data.frame(unique(SN$MOTIVO))
i <- row(SN_MOTIVO[1])
for (a in i){
  SN_MOTIVO$Desc[a] <- a
}

names(SN_MOTIVO) <- c("MOTIVO","COD_MOTIVO")
SN1 <- inner_join(SN,SN_MOTIVO,by="MOTIVO")

arvore <- rpart(SIT_CADASTRAL ~ QTD + TEMPO_FUNCIONAMENTO ,
             data = SN1, method = "class")

#arvore <- rpart(COD_MOTIVO ~ TEMPO_FUNCIONAMENTO + QTD,
#              data = SN1, method = "class")

printcp(arvore)          # Exibe os resultados
plotcp(arvore)           # visualizar os resultados de cross-validation
summary(arvore)          # detalhamento da montagem da árvore

rpart.plot(prune(arvore, cp = 0.01000000),
         tweak = 1.2,
         type = 5,
         branch = .9,
         yesno = F,
         extra = 2, under = T,
         legend.x = NA,
         gap = 1,space = .5,
         shadow.col = "GRAY",
         main = "Empresas do Simples Nacional\nPeríodo: 2005 a 2019\n")



rm(a,i,SN1,arvore,SN_MOTIVO)
```

```
> printcp(arvore)           # Exibe os resultados

Classification tree:
rpart(formula = SIT_CADASTRAL ~ QTD + TEMPO_FUNCIONAMENTO, data = SN1,
    method = "class")

Variables actually used in tree construction:
[1] QTD                 TEMPO_FUNCIONAMENTO

Root node error: 2001/5456 = 0.36675

n= 5456

        CP nsplit rel error  xerror     xstd
1 0.193903      0   1.00000 1.00000 0.017789
2 0.053223      1   0.80610 0.80660 0.016848
3 0.020990      3   0.69965 0.71564 0.016241
4 0.010000      4   0.67866 0.69315 0.016073
```



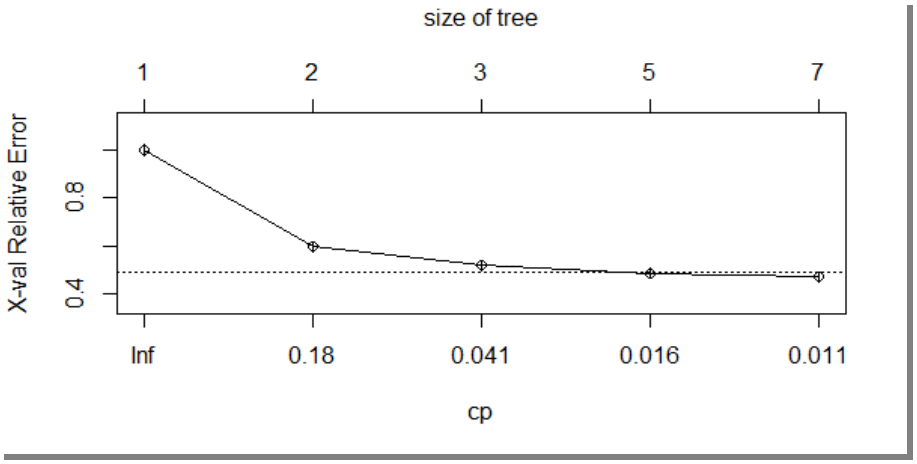


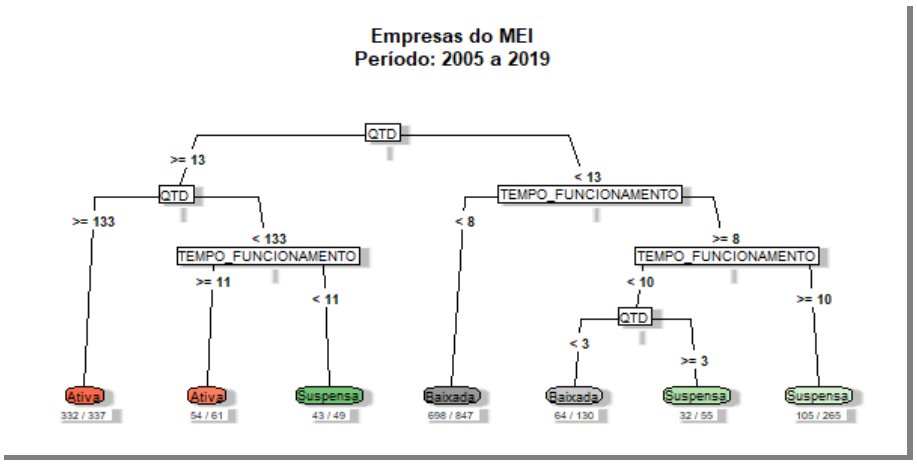
E por último o script, assim como os resultados, para as Empresas optantes pelo MEI está demonstrado abaixo.

```
################################################################################
################ PARA O MEI
################################################################################
```

```
MEI_MOTIVO <- as.data.frame(unique(MEI$MOTIVO))
i <- row(MEI_MOTIVO[1])
for (a in i){
  MEI_MOTIVO$Desc[a] <- a
}

names(MEI_MOTIVO) <- c("MOTIVO","COD_MOTIVO")
MEI1 <- inner_join(MEI,MEI_MOTIVO,by="MOTIVO")

arvore <- rpart(SIT_CADASTRAL ~ QTD + TEMPO_FUNCIONAMENTO ,
             data = MEI1, method = "class")

#arvore <- rpart(COD_MOTIVO ~ SIT_CADASTRAL + TEMPO_FUNCIONAMENTO + QTD,
#              data = MEI1, method = "class")

printcp(arvore)          # Exibe os resultados
plotcp(arvore)           # visualizar os resultados de cross-validation
summary(arvore)          # detalhamento da montagem da árvore

rpart.plot(prune(arvore, cp = 0.010000),
         tweak = 1.2,
         type = 5,
         branch = .9,
         yesno = F,
         extra = 2, under = T,
         legend.x = NA,
         gap = 1,space = .5,
         shadow.col = "GRAY",
         main = "Empresas do MEI\nPeríodo: 2005 a 2019\n")

rm(a,i,MEI1,arvore,MEI_MOTIVO)

################################################################################
```

```
> printcp(arvore)           # Exibe os resultados

Classification tree:
rpart(formula = SIT_CADASTRAL ~ QTD + TEMPO_FUNCIONAMENTO, data = MEI1,
    method = "class")

Variables actually used in tree construction:
[1] QTD                 TEMPO_FUNCIONAMENTO

Root node error: 935/1744 = 0.53612

n= 1744

        CP nsplit rel error  xerror     xstd
1 0.410695      0   1.00000 1.00000 0.022274
2 0.074866      1   0.58930 0.59786 0.020844
3 0.022995      2   0.51444 0.51872 0.020012
4 0.011765      4   0.46845 0.48770 0.019627
5 0.010000      6   0.44492 0.47059 0.019399
```





Cabe ressaltar que os modelos de árvore de decisão devem ser verificados de acordo com os resultados obtidos. O pacote “rpart” fornece ferramentas para verificar os erros do nó raiz, que é a porcentagem de registros classificados corretamente no primeiro nó de divisão (raiz).

Ouras medidas de desempenho devem ser observadas: Root Node Error x Rel Error é a taxa de erro de re-substituição (a taxa de erro calculada na amostra de treinamento); Erro de nó raiz x X Erro é a taxa de erro de validação cruzada, que é uma medida mais objetiva de precisão preditiva, e n é o número de registros usados para construir a árvore.

Portanto, como este é um modelo supervisionado, deve-se verificar qual o melhor tratamento a ser dado, para que se obtenha o melhor resultado.

O segundo modelo que iremos apresentar pretende avaliar os motivos das situações cadastrais das empresas, permitindo uma atuação, ou pelo menos um direcionamento, das ações preventivas que permitissem aumentar o tempo de funcionamento, salutar, da empresa.

Para isso, utilizaremos a abordagem denominada Mineração de Texto, incluindo uma análise de sentimentos, cuja base de polaridade (Positiva e negativa) foi elaborada com base nos DataSet's estudados.

Para o desenvolvimento do algoritmo de treinamento utilizamos o pacote "rtweet", "wordcloud" e "tidytext" do R, cujas funcionalidades permitem o tratamento adequado na obtenção do modelo pretendido.

Apresentamos o script utilizado na modelagem, iniciando com o carregamento das bibliotecas necessárias (library's no R).

```
################################################################################
############## BIBLIOTECAS PARA TEXT MINING
################################################################################
library(rtweet)
library(wordcloud)
library(tidytext)

################################################################################
############## ANÁLISE DOS MOTIVOS DA EXTINÇÃO COM TEXT MINING
################################################################################

base_polaridade <- read.csv2(file=paste0(base_path,"base_polaridade.csv"),sep = ",",fileEncoding="UTF-8")
sw <- read.csv2(file=paste0(base_path,"sw.csv"),sep = ";",fileEncoding="UTF-8")
```

```
################################################################################
##############        TEXT MINING - LUCRO REAL/PRESUMIDO
################################################################################

analise_mensagem <- LR %>%
  unnest_tokens(palavra, MOTIVO,to_lower = TRUE) %>%
  group_by(ANO_ABERTURA , palavra) %>%
  summarise(n = sum(QTD))

total_palavras <- analise_mensagem %>%
  group_by(ANO_ABERTURA) %>%
  summarize(total=sum(n))

analise_mensagem <- inner_join(analise_mensagem, total_palavras)

analise_mensagem <- analise_mensagem %>%
  bind_tf_idf(palavra, ANO_ABERTURA, n) #%>%
  #filter(tf > 0.0001)

stop_words_grupo  <-  unique(c(unique(analise_mensagem$palavra[analise_mensagem$tf  ==
0.00]))) #, stopwords::stopwords("pt")))
stop_words_grupo                                                                    <-
c(stop_words_grupo,c("de","a","das","localizacao","pela","do","rfb","convenente","nao","pelo"
,"na","2009","em","e","da","me","dado","2006","as","por","pedido"))

analise_mensagem <- analise_mensagem %>%
  anti_join(data_frame(palavra = stop_words_grupo))

p <- analise_mensagem %>%
#  anti_join(data_frame(palavra = stop_words_grupo1)) %>%
  group_by(palavra)%>%
  summarise(
    n = sum(n)
  ) %>%
  filter(n >= 300) %>%
  ungroup() %>%
  mutate(palavra = reorder(palavra, n)) %>%
  ggplot(aes(x=palavra, y=n)) +
```

```
  geom_segment( aes(x=palavra, xend=palavra, y=0, yend=n ) ) +
  geom_point(color = "orange" ) +
  theme_light(base_size = 12, base_family = "") +
  coord_flip() +
  theme(
    legend.position="none",
    panel.grid.major.y = element_blank(),
    axis.ticks.length = unit(.99, "cm"),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  )

print(p)
p + labs(title = "Text-Minig - Lucro Real/Presumido\nPeríodo: 2005 a 2019") + xlab("Ocorrências") + ylab("contagem")

analise_mensagem %>%
#  anti_join(data_frame(palavra = stop_words_grupo1)) %>%
  group_by(palavra)%>%
  summarise(
    n = sum(n)
  ) %>%
  with(wordcloud(palavra,n,max.words = 15, colors=brewer.pal(6,"Dark2"),random.order=FALSE))

analise_mensagem %>%
  filter(n > 150) %>%
  mutate(n = ifelse(palavra == "negative", -n, n)) %>%
  mutate(word = reorder(ANO_ABERTURA, n)) %>%
  ggplot(aes(ANO_ABERTURA, n, fill = palavra)) +
  geom_col() +
  coord_flip() +
  labs(y = "Contribuição de cada ocorrência \n Empresas Lucro Real/Presumido - Período: 2005 a 2019")

pal_tmp <- analise_mensagem[,1:3]
pal_tmp <- inner_join(pal_tmp,base_polaridade,"palavra")
pal_tmp <- pal_tmp[pal_tmp$n > 50000,]
```

```
tb_bars <- pal_tmp %>%
  mutate(n = polaridade * n) %>%
  group_by(polaridade) %>%
  ungroup()

p <- ggplot(data = tb_bars,
      mapping = aes(x = reorder(palavra, n),
                    y = n,
                    fill = n)) +
  geom_col(color = "black") +
  scale_fill_distiller(palette = "RdBu", direction = 1) +
  coord_flip() +
  theme_light() +
  #theme(legend.position = c(0.95, 0.5),
  #     legend.justification = c(1, 0.5)) +
  labs(y = "Frequência de ocorrência",
      x = "Motivo",
      fill = "Frequência de\nocorrência")

p
p + labs(title = "Análise do motivo de encerramento", subtitle = "Empresas do Lucro Real/Presumido - Período: 2005 - 2019")

rm(p,analise_mensagem,stop_words_grupo,total_palavras,tb_bars,n_words,pal_tmp)

################################################################################
##############        TEXT MINING - SIMPLES NACIONAL
################################################################################

analise_mensagem <- SN %>%
  unnest_tokens(palavra, MOTIVO,to_lower = TRUE) %>%
  group_by(ANO_ABERTURA , palavra) %>%
  summarise(n = sum(QTD))

total_palavras <- analise_mensagem %>%
  group_by(ANO_ABERTURA) %>%
  summarize(total=sum(n))
```

```
analise_mensagem <- inner_join(analise_mensagem, total_palavras)

analise_mensagem <- analise_mensagem %>%
  bind_tf_idf(palavra, ANO_ABERTURA, n) #%>%
#  filter(tf > 0.01)

stop_words_grupo  <-  unique(c(unique(analise_mensagem$palavra[analise_mensagem$tf  ==
0.00]))) #, stopwords::stopwords("pt")))
stop_words_grupo                                                                    <-
c(stop_words_grupo,c("de","a","das","localizacao","pela","do","rfb","convenente","nao","pelo"
,"na","2009","em","e","da","me","dado","2006","as","por","pedido"))

analise_mensagem <- analise_mensagem %>%
  anti_join(data_frame(palavra = stop_words_grupo))

p <- analise_mensagem %>%
#  anti_join(data_frame(palavra = stop_words_grupo)) %>%
  group_by(palavra)%>%
  summarise(
    n = sum(n)
  ) %>%
  filter(n >= 45) %>%
  ungroup() %>%
  mutate(palavra = reorder(palavra, n)) %>%
  ggplot(aes(x=palavra, y=n)) +
  geom_segment( aes(x=palavra, xend=palavra, y=0, yend=n ) ) +
  geom_point(color = "orange" ) +
  theme_light(base_size = 12, base_family = "") +
  coord_flip() +
  theme(
    legend.position="none",
    panel.grid.major.y = element_blank(),
    axis.ticks.length = unit(.99, "cm"),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  )
```

```
print(p)
p + labs(title = "Text-Minig - Simples Nacional\nPeríodo: 2005 a 2019")  + xlab("Ocorrências") +
ylab("contagem")

analise_mensagem %>%
#  anti_join(data_frame(palavra = stop_words_grupo)) %>%
 group_by(palavra)%>%
 summarise(
   n = sum(n)
 ) %>%
 with(wordcloud(palavra,n,max.words                    =                    15,
colors=brewer.pal(6,"Dark2"),random.order=FALSE))

analise_mensagem %>%
 filter(n > 150) %>%
 mutate(n = ifelse(palavra == "negative", -n, n)) %>%
 mutate(word = reorder(ANO_ABERTURA, n)) %>%
 ggplot(aes(ANO_ABERTURA, n, fill = palavra)) +
 geom_col() +
 coord_flip() +
 labs(y = "Contribuição de cada ocorrência \n Empresas do Simples Nacional - Período: 2005 a
2019")

pal_tmp <- analise_mensagem[,1:3]
pal_tmp <- inner_join(pal_tmp,base_polaridade,"palavra")
pal_tmp <- pal_tmp[pal_tmp$n > 500,]

tb_bars <- pal_tmp %>%
 mutate(n = polaridade * n) %>%
 group_by(polaridade) %>%
#  top_n(n, n = n_words) %>%
#  top_n(n, n = n_words) %>%
 ungroup()

p <- ggplot(data = tb_bars,
      mapping = aes(x = reorder(palavra, n),
                    y = n,
                    fill = n)) +
```

```
  geom_col(color = "black") +
  scale_fill_distiller(palette = "RdBu", direction = 1) +
  coord_flip() +
  theme_light() +
#  theme(legend.position = c(0.95, 0.5),
#        legend.justification = c(1, 0.5)) +
  labs(y = "Frequência de ocorrência",
       x = "Motivo",
       fill = "Frequência de\nocorrência")
p
p + labs(title = "Análise do motivo de encerramento", subtitle = "Empresas do Simples Nacional - Período: 2005 - 2019")

rm(p,analise_mensagem,stop_words_grupo,total_palavras,tb_bars,n_words,pal_tmp)

################################################################################
###############        TEXT MINING - MEI
################################################################################

analise_mensagem <- MEI %>%
  unnest_tokens(palavra, MOTIVO,to_lower = TRUE) %>%
  group_by(ANO_ABERTURA , palavra) %>%
  summarise(n = sum(QTD))

total_palavras <- analise_mensagem %>%
  group_by(ANO_ABERTURA) %>%
  summarize(total=sum(n))

analise_mensagem <- inner_join(analise_mensagem, total_palavras)

analise_mensagem <- analise_mensagem %>%
  bind_tf_idf(palavra, ANO_ABERTURA, n) # %>%
#  filter(tf > 0.0001)

stop_words_grupo  <-  unique(c(unique(analise_mensagem$palavra[analise_mensagem$tf  ==  0.00]))) #, stopwords::stopwords("pt")))
```

```
stop_words_grupo <-
c(stop_words_grupo,c("de","a","das","localizacao","pela","do","rfb","convenente","nao","pelo"
,"na","2009","em","e","da","me","dado","2006","as","por","pedido"))

analise_mensagem <- analise_mensagem %>%
 anti_join(data_frame(palavra = stop_words_grupo))

p <- analise_mensagem %>%
#  anti_join(data_frame(palavra = stop_words_grupo)) %>%
 group_by(palavra)%>%
 summarise(
  n = sum(n)
 ) %>%
 filter(n >= 150) %>%
 ungroup() %>%
 mutate(palavra = reorder(palavra, n)) %>%
 ggplot(aes(x=palavra, y=n)) +
 geom_segment( aes(x=palavra, xend=palavra, y=0, yend=n ) ) +
 geom_point(color = "orange" ) +
 theme_light(base_size = 12, base_family = "") +
 coord_flip() +
 theme(
  legend.position="none",
  panel.grid.major.y = element_blank(),
  axis.ticks.length = unit(.99, "cm"),
  panel.border = element_blank(),
  axis.ticks.y = element_blank()
 )

print(p)
p + labs(title = "Text-Minig - MEI\nPeríodo: 2005 a 2019")  + xlab("Ocorrências") +
ylab("contagem")

analise_mensagem %>%
#  anti_join(data_frame(palavra = stop_words_grupo)) %>%
 group_by(palavra)%>%
 summarise(
  n = sum(n)
```

```
 ) %>%
 with(wordcloud(palavra,n,max.words = 15,
colors=brewer.pal(6,"Dark2"),random.order=FALSE))

analise_mensagem %>%
 filter(n > 150) %>%
 mutate(n = ifelse(palavra == "negative", -n, n)) %>%
 mutate(word = reorder(ANO_ABERTURA, n)) %>%
 ggplot(aes(ANO_ABERTURA, n, fill = palavra)) +
 geom_col() +
 coord_flip() +
 labs(y = "Contribuição de cada ocorrência \n Empresas do MEI - Período: 2005 a 2019")

pal_tmp <- analise_mensagem[,1:3]
pal_tmp <- inner_join(pal_tmp,base_polaridade,"palavra")
pal_tmp <- pal_tmp[pal_tmp$n > 50,]

tb_bars <- pal_tmp %>%
 mutate(n = polaridade * n) %>%
 group_by(polaridade) %>%
 #  top_n(n, n = n_words) %>%
 #  top_n(n, n = n_words) %>%
 ungroup()

p <- ggplot(data = tb_bars,
        mapping = aes(x = reorder(palavra, n),
                      y = n,
                      fill = n)) +
 geom_col(color = "black") +
 scale_fill_distiller(palette = "RdBu", direction = 1) +
 coord_flip() +
 theme_light() +
 #  theme(legend.position = c(0.95, 0.5),
 #        legend.justification = c(1, 0.5)) +
 labs(y = "Frequência de ocorrência",
     x = "Motivo",
     fill = "Frequência de\nocorrência")
p
```

```
p + labs(title = "Análise do motivo de encerramento", subtitle = "Empresas do MEI - Período: 2005 - 2019")

rm(p,analise_mensagem,stop_words_grupo,total_palavras,tb_bars,n_words,pal_tmp)
```

O script inicia carregando uma base de polaridade, onde estão classificadas todas as palavras que compõe os motivos das situações cadastrais. Um pequeno exemplo desta base está ilustrada abaixo.

palavra	polaridade	tipo	sentimento
54	-1	adjetivo	negativo
123	1	adjetivo	positivo
11941	-1	adjetivo	negativo
11941	-1	adjetivo	negativo

Posteriormente, foram incluídas as denominadas StopWords, que conceitualmente são palavras consideradas irrelevantes, e que devem ser descartadas do texto que se está analisando.

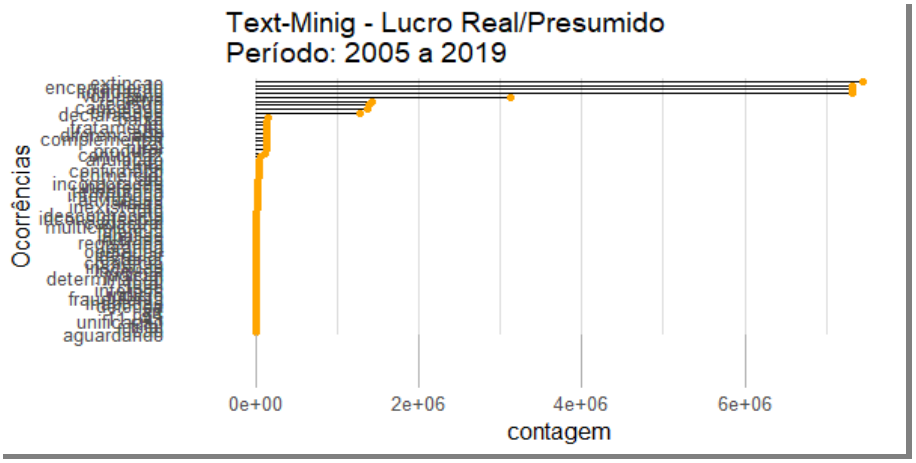
Houve a necessidade de efetuarmos um tratamento nos dados de leitura antes de prosseguirmos com o modelo.

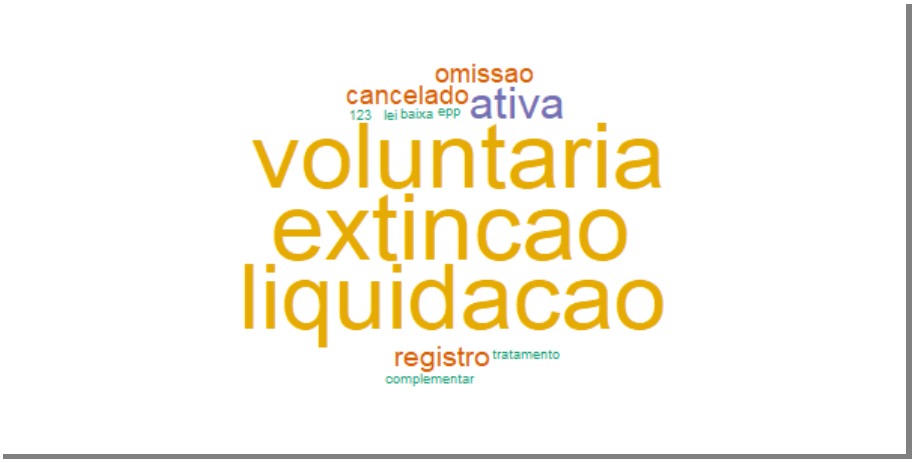
Os Dataset's estão agrupados por alguns atributos: UF, Ano_Abertura, Ano_Baixa, Status_Cadastro e Motivo_Baixa. E a quantidade de estabelecimentos se refere a cada uma destas composições (tuplas). Então, por exemplo, se tivermos 500 estabelecimentos para a tupla: 'SP',2005,2010,"Ativo","ativa", onde o atributo motivo da situação cadastral está com o valor "Ativa", teremos que introduzir na base de classificação a palavra – "Ativa", para esta tupla 500 vezes.

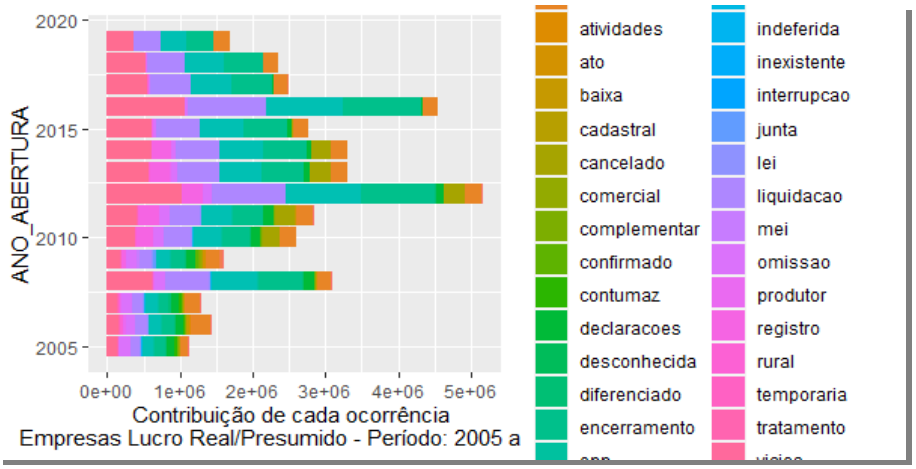
Se não fizermos este tratamento uma tupla com 100 estabelecimentos (QTD), terá o mesmo número de ocorrências que uma tupla com 500 estabelecimentos.

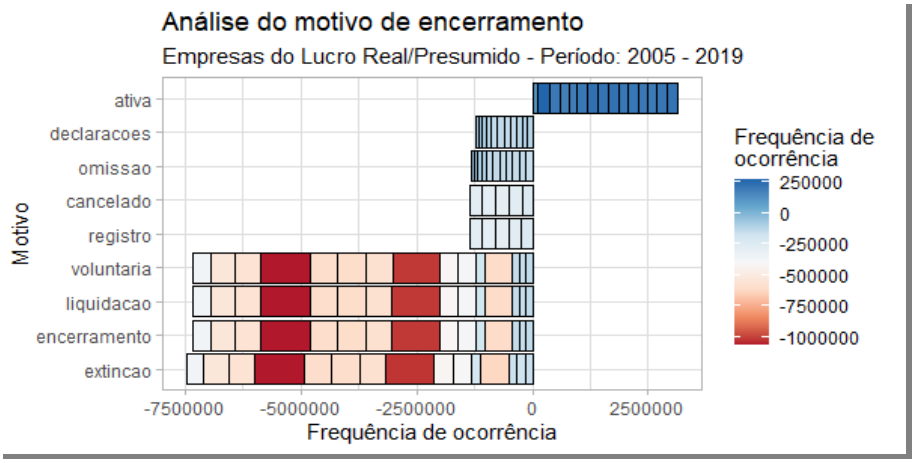
Finalizada a montagem correta da base de pesquisa, com a quebra das palavras que compõe o atributo Motivo, efetuamos a sua classificação por intermédio da base de polaridade.

Os resultados para as empresa do Lucro Real/Presumido estão demonstrados abaixo:

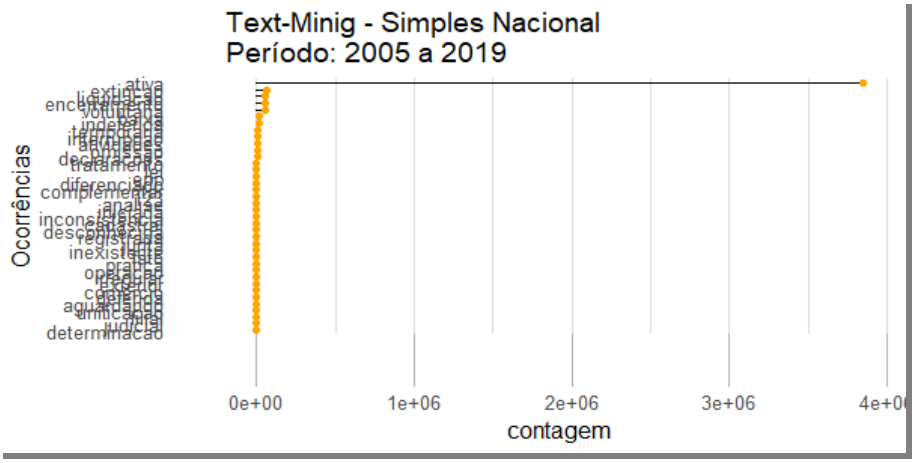

Text-Minig - Lucro Real/Presumido
Período: 2005 a 2019
Ocorrências
0e+00
2e+06
4e+06
6e+06
contagem

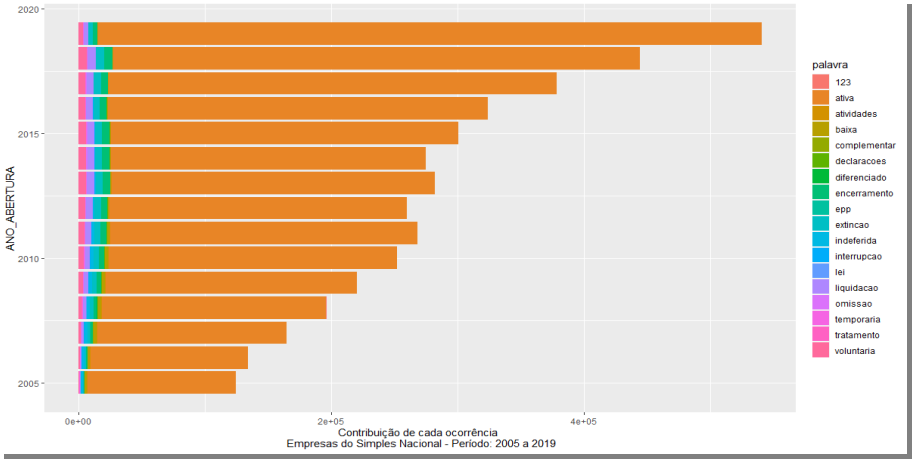

omissao
cancelado
ativa
123
lei
baixa
epp
voluntaria
extincao
liquidacao
registro
tratamento
complementar

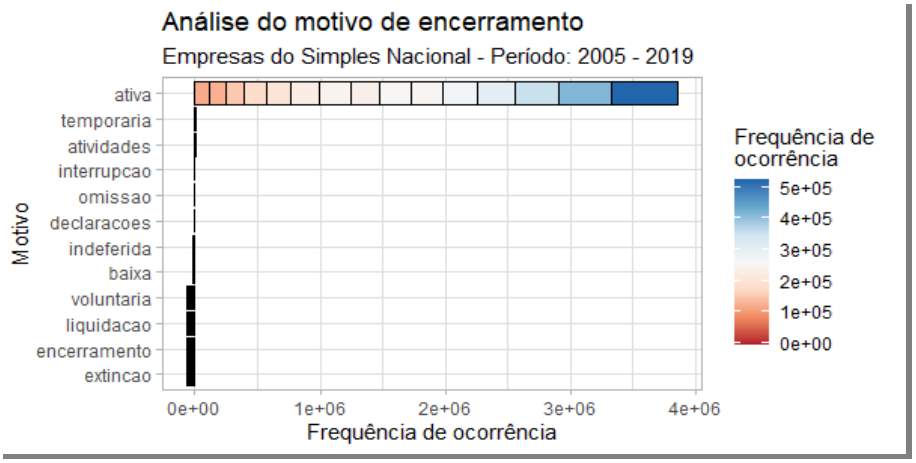

2020
2015
2010
2005
ANO_ABERTURA
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Contribuição de cada ocorrência
Empresas Lucro Real/Presumido - Período: 2005 a
atividades
ato
baixa
cadastral
cancelado
comercial
complementar
confirmado
contumaz
declaracoes
desconhecida
diferenciado
encerramento
indeferida
inexistente
interrupcao
junta
lei
liquidacao
mei
omissao
produtor
registro
rural
temporaria
tratamento



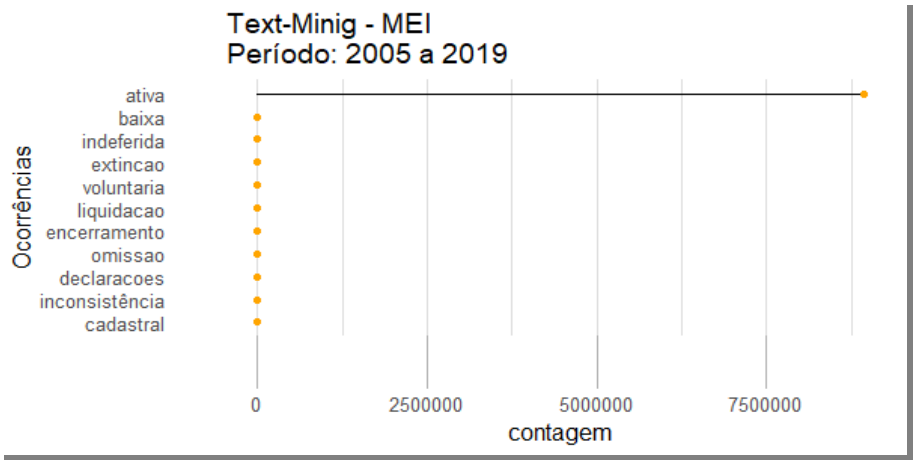
Os resultados para as empresa do Simples Nacional estão demonstrados abaixo:



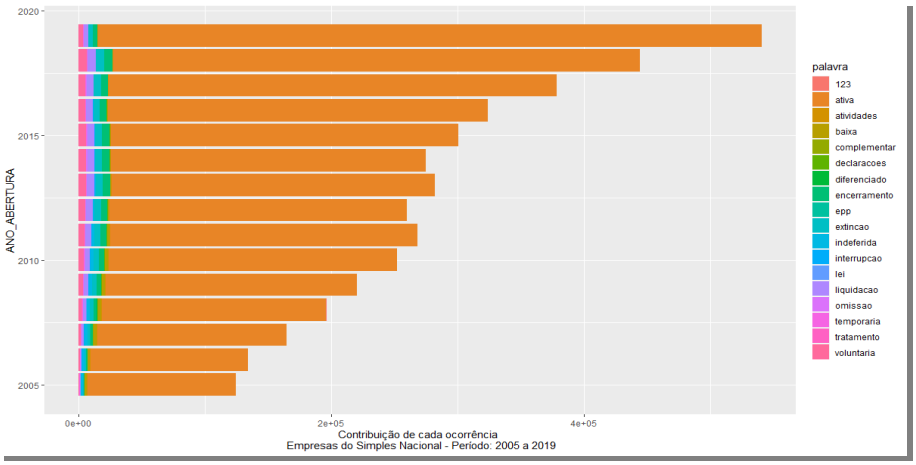


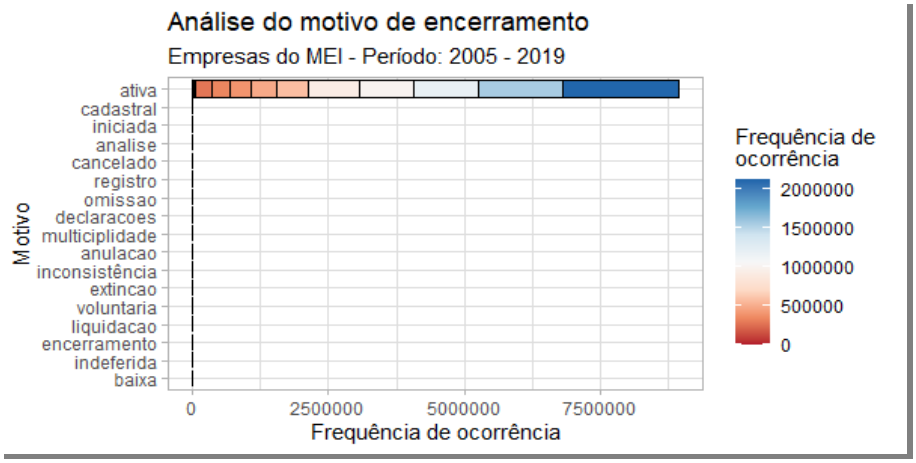


Os resultados para as empresa do MEI estão demonstrados abaixo:




inconsistência
desconhecida
declaracoes
extincao
omissao
indeferida
ativa
anulacao
baixa voluntaria
encerramento
analise
liquidacao
cadastral
iniciada


2020
2015
2010
2005
ANO_ABERTURA
0e+00
2e+05
4e+05
Contribuição de cada ocorrência
Empresas do Simples Nacional - Período: 2005 a 2019
palavra
123
ativa
atividades
baixa
complementar
declaracoes
diferenciado
encerramento
epp
extincao
indeferida
interrupcao
lei
liquidacao
omissao
temporaria
tratamento
voluntaria


Análise do motivo de encerramento
Empresas do MEI - Período: 2005 - 2019
Motivo
ativa
cadastral
iniciada
analise
cancelado
registro
omissao
declaracoes
multiciplidade
anulacao
inconsistência
extincao
voluntaria
liquidacao
encerramento
indeferida
baixa
0
2500000
5000000
7500000
Frequência de ocorrência
Frequência de ocorrência
2000000
1500000
1000000
500000
0

6. Apresentação dos Resultados

Iniciaremos a apresentação dos resultados seguindo o modelo de workflow proposto por Vasandani (clique aqui)

Título: **O REGIME DE TRIBUTAÇÃO NO CICLO DE VIDA DAS EMPRESAS**		
Definição do problema	**Resultados e previsões**	**Aquisição de dados**
Verificar a influência do regime de tributação no ciclo de vida das empresas.	Objetivamos poder verificar se o regime de tributação escolhido no momento da criação de uma empresa pode influenciar no seu ciclo de vida. Desta forma, poderemos agir de forma preventiva, evitando ou pelo menos minimizando os efeitos da influência do regime escolhido.	Os dados utilizados no presente trabalho foram obtidos de uma base governamental – Cadastro Nacional das Pessoas Jurídicas – CNPJ. O período utilizado na Análise Exploratória e treinamento dos modelos correspondem aos anos de 2005 a 2019. Para os dados de teste dos modelos treinados, foi utilizado o ano de 2020.
Modelagem	**Avaliação do modelo**	**Preparação dos dados**
São realizadas análises exploratórias utilizando scrips em R, assim como modelos de classificação, e identificação dos motivos que levaram as empresas ao fim de	Assim que obtivemos os modelos de treinamento, estes foram verificados com base nos resultados apresentados, verificando-se comparativamente os dois modelos obtidos.	Os dados foram preparados, inicialmente, na extração do DW, e posteriormente alguns campos foram ajustados pra permitirem a

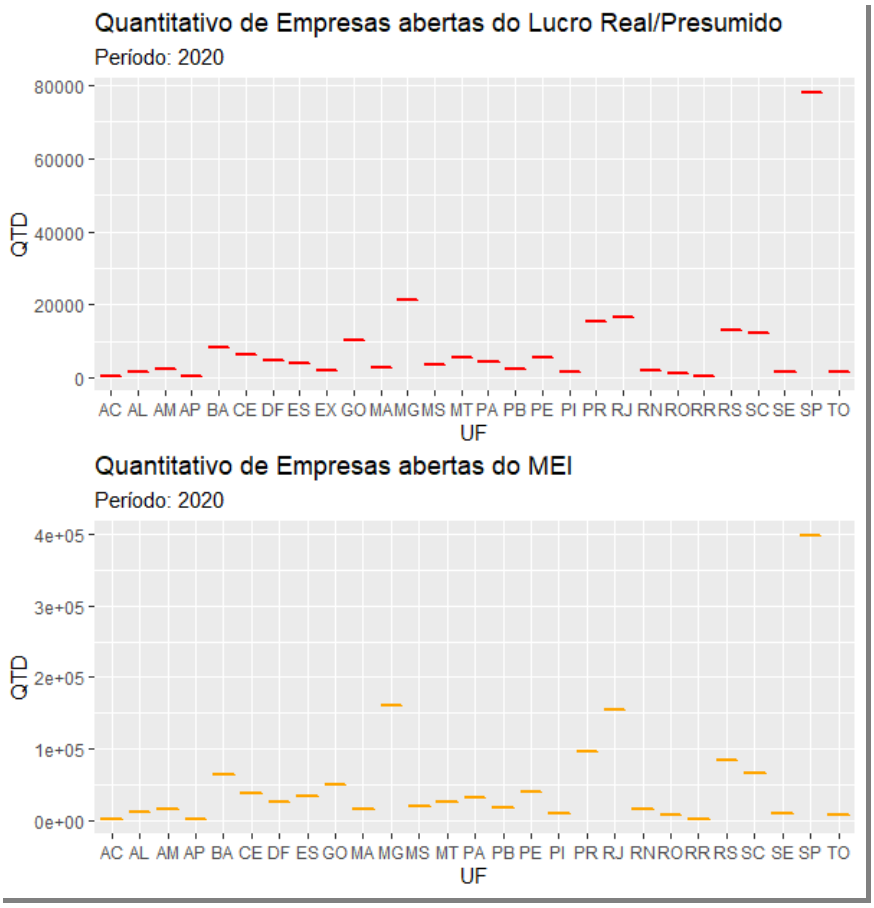
suas atividades.		modelagem correta.

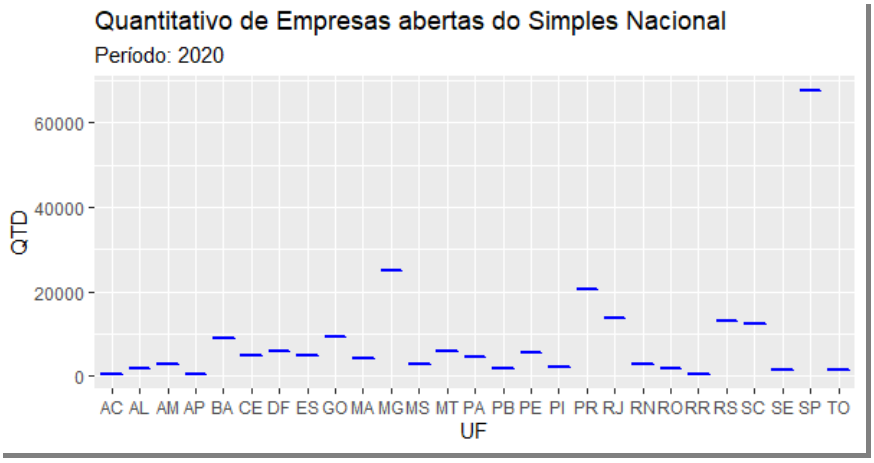
Para verificação dos resultados efetuaremos os modelos utilizando os Data-Set's com os registros das empresas com data de abertura igual ao ano corrente – 2020. Apesar de ser um ano, completamente, atípico esperamos que os modelos possam nos permitir avaliar a situação das empresas.

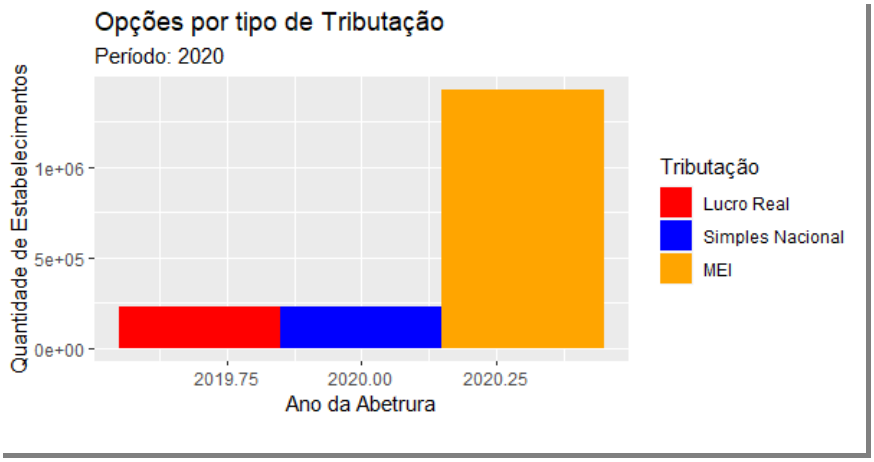
Ressaltamos que os algoritmos foram preparados para executarem, com uma pequena alteração do Dataset a ser estudado. Portanto, não é necessário alterar-mos, o algoritmo para obtermos os resultados do modelo.

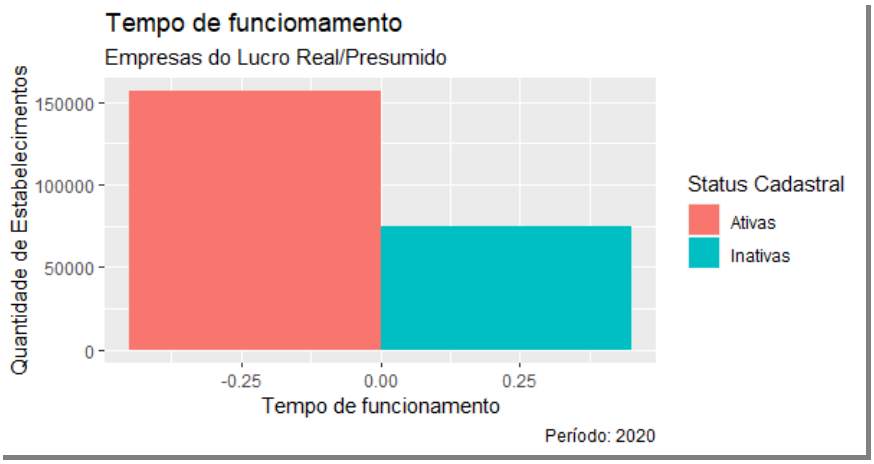
Os Dataset's a serem utilizados nos testes são: LR_2020 (empresas do Lucro Real/Presumido abertas em 2020), SN_2020 (empresas do Simples Nacional abertas em 2020) e MEI_2020 (empresas do MEI abertas em 2020).

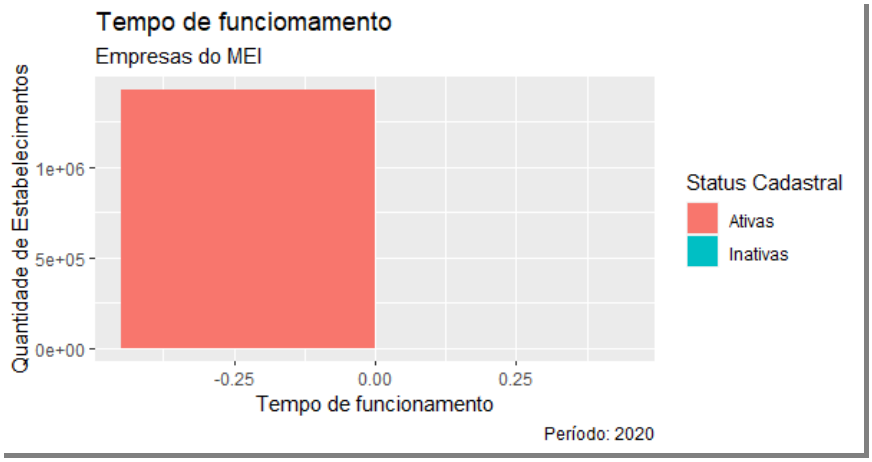
Inicialmente mostraremos os resultados gerais da base de testes, e posteriormente os resultados dos modelos escolhidos.

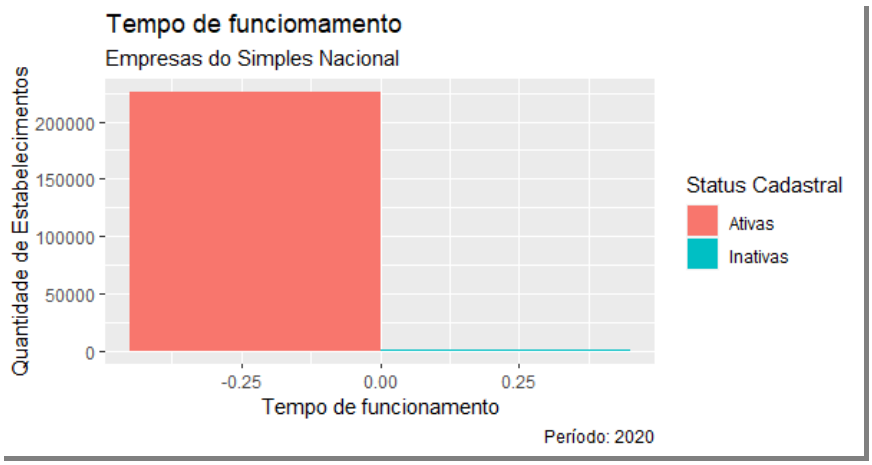



Quantitativo de Empresas abertas do Simples Nacional
Período: 2020
QTD
60000
40000
20000
0
AC AL AM AP BA CE DF ES GO MA MG MS MT PA PB PE PI PR RJ RN RO RR RS SC SE SP TO
UF


Opções por tipo de Tributação
Período: 2020
Quantidade de Estabelecimentos
1e+06
5e+05
0e+00
2019.75
2020.00
2020.25
Ano da Abetrura
Tributação
Lucro Real
Simples Nacional
MEI

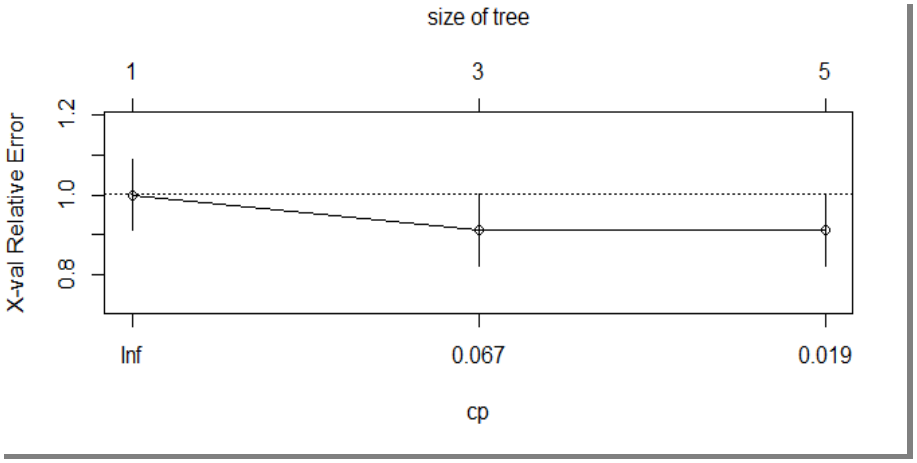

Tempo de funciomamento
Empresas do Lucro Real/Presumido
Quantidade de Estabelecimentos
150000
100000
50000
0
-0.25
0.00
0.25
Tempo de funcionamento
Período: 2020
Status Cadastral
Ativas
Inativas

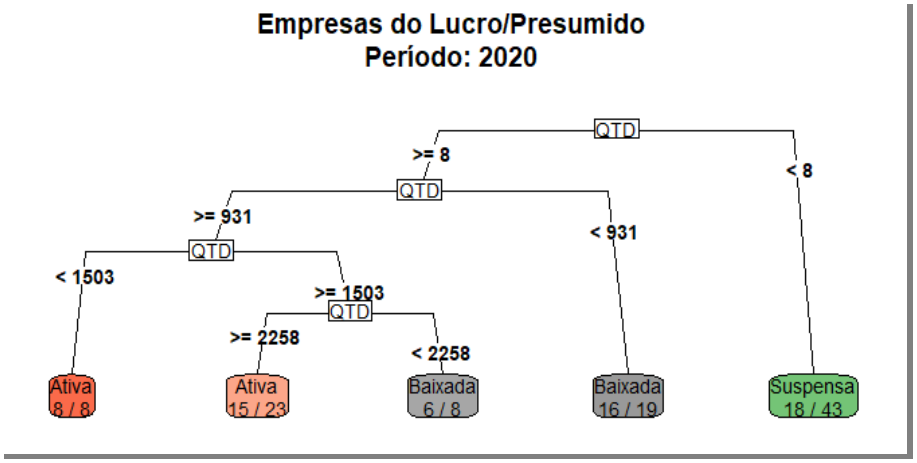


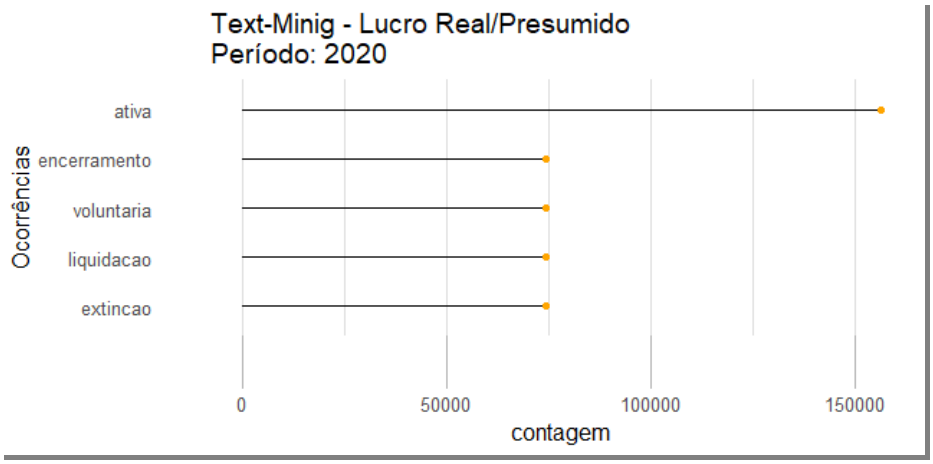


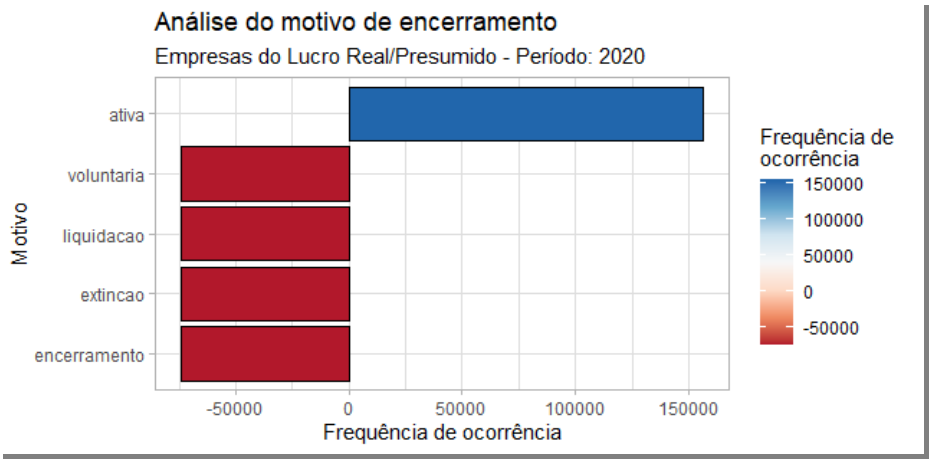
Os resultados após a aplicação dos testes nas empresas optantes pelo Lucro Real/Presumido estão dispostos abaixo.

```
> printcp(arvore)

Classification tree:
rpart(formula = SIT_CADASTRAL ~ TEMPO_FUNCIONAMENTO + QTD, data = LR1,
    method = "class")

Variables actually used in tree construction:
[1] QTD

Root node error: 56/101 = 0.55446

n= 101

        CP nsplit rel error  xerror     xstd
1 0.125000      0   1.00000 1.00000 0.089197
2 0.035714      2   0.75000 0.91071 0.089727
3 0.010000      4   0.67857 0.91071 0.089727
>
```


size of tree
1
3
5
X-val Relative Error
0.8
1.0
1.2
Inf
0.067
0.019
cp

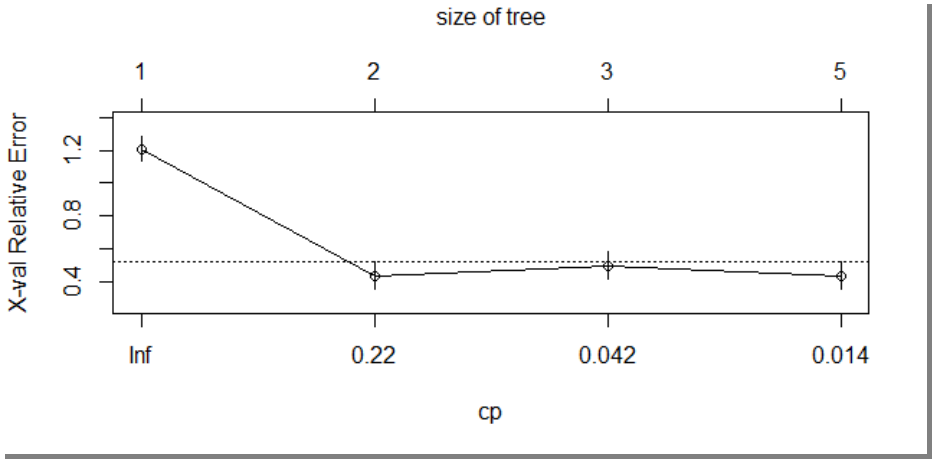

Empresas do Lucro/Presumido
Período: 2020
QTD
>= 8
< 8
QTD
>= 931
< 931
QTD
< 1503
>= 1503
QTD
>= 2258
< 2258
Ativa
8 / 8
Ativa
15 / 23
Baixada
6 / 8
Baixada
16 / 19
Suspensa
18 / 43

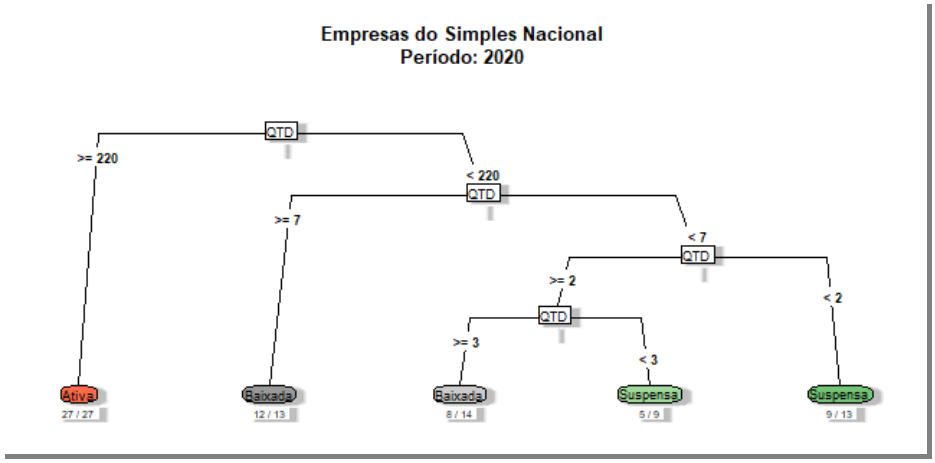

Text-Minig - Lucro Real/Presumido
Período: 2020
ativa
encerramento
voluntaria
liquidacao
extincao
Ocorrências
0
50000
100000
150000
contagem

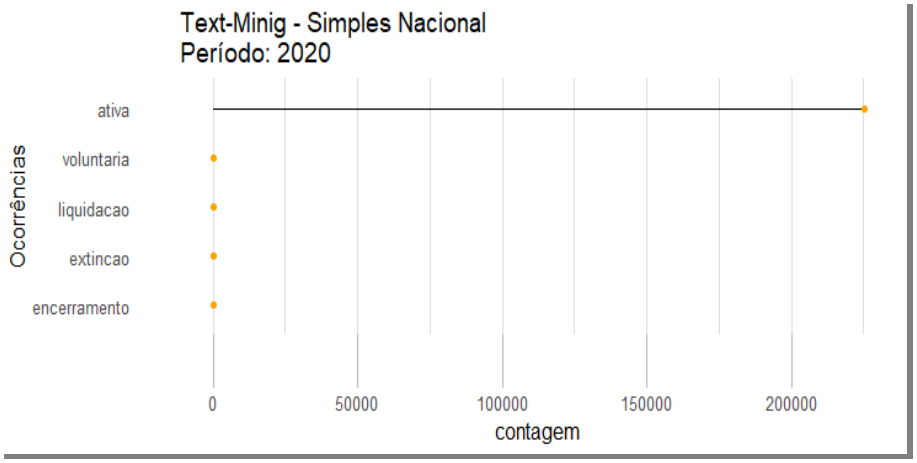


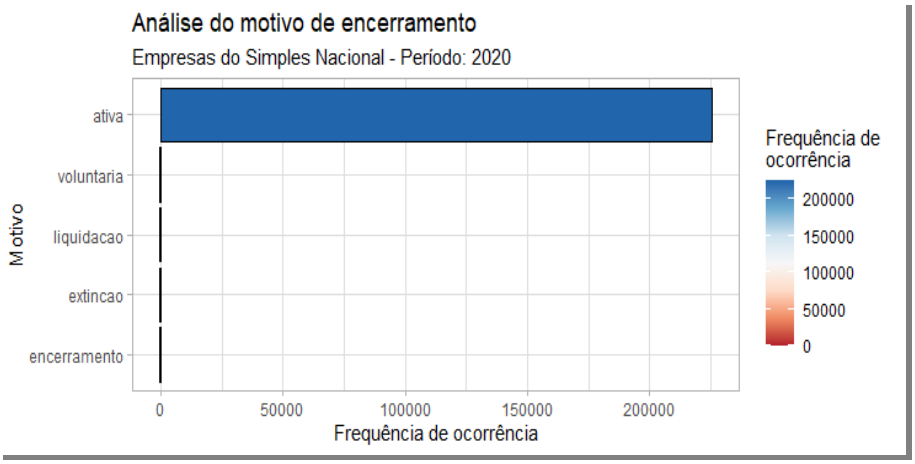
Os resultados após a aplicação dos testes nas empresas optantes pelo Simples Nacional estão dispostos abaixo.

```
> printcp(arvore)           # Ex

Classification tree:
rpart(formula = SIT_CADASTRAL ~ QTD + TEMPO_FUNCIONAMENTO, data = SN1,
    method = "class")

Variables actually used in tree construction:
[1] QTD

Root node error: 48/76 = 0.63158

n= 76

        CP nsplit rel error xerror     xstd
1 0.562500      0   1.00000 1.2083 0.077215
2 0.083333      1   0.43750 0.4375 0.081216
3 0.020833      2   0.35417 0.5000 0.084423
4 0.010000      4   0.31250 0.4375 0.081216
>
```


size of tree
1
2
3
5
X-val Relative Error
0.4
0.8
1.2
Inf
0.22
0.042
0.014
cp

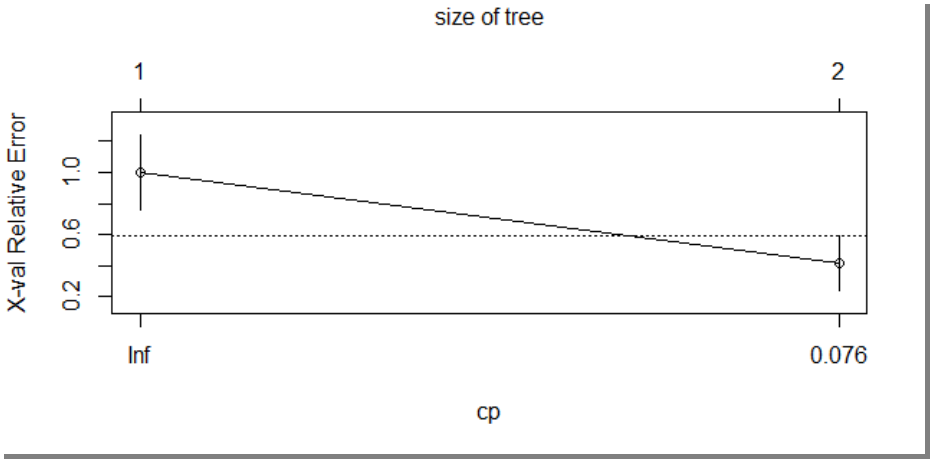

Empresas do Simples Nacional
Período: 2020
QTD
>= 220
< 220
QTD
>= 7
< 7
QTD
>= 2
QTD
< 2
>= 3
< 3
Ativa
27 / 27
Baixada
12 / 13
Baixada
8 / 14
Suspensa
5 / 9
Suspensa
9 / 13

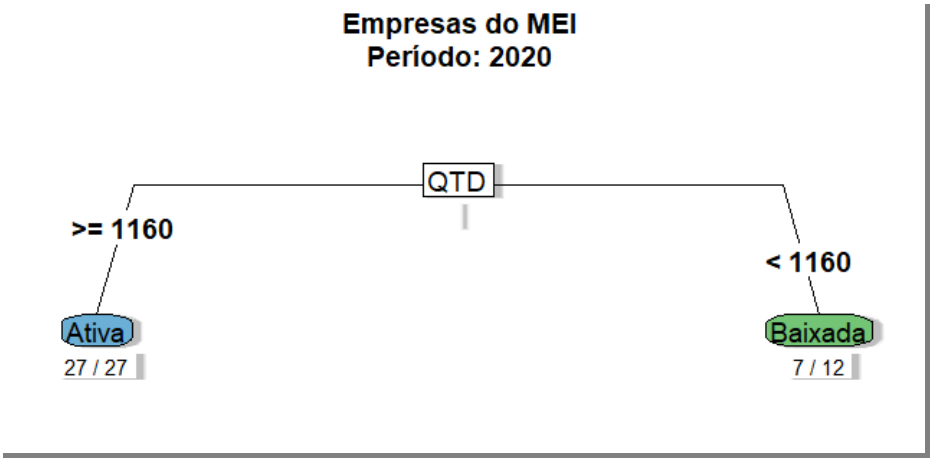

Text-Minig - Simples Nacional
Período: 2020
Ocorrências
ativa
voluntaria
liquidacao
extincao
encerramento
0
50000
100000
150000
200000
contagem

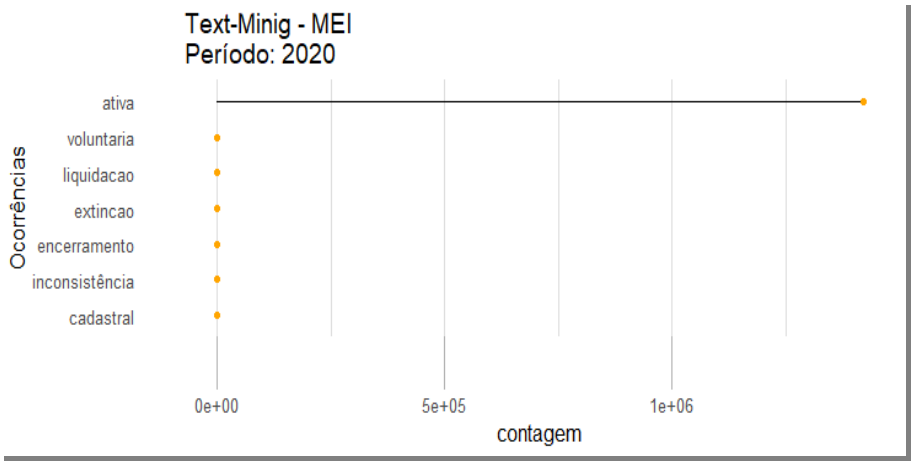


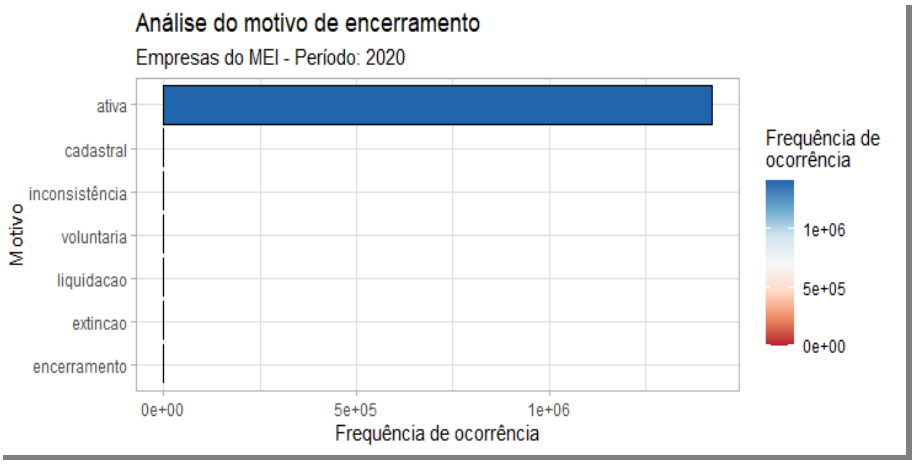
Os resultados após a aplicação dos testes nas empresas optantes pelo MEI estão dispostos abaixo.

```
> printcp(arvore)          #

Classification tree:
rpart(formula = SIT_CADASTRAL ~ QTD + TEMPO_FUNCIONAMENTO, data = MEI1,
    method = "class")

Variables actually used in tree construction:
[1] QTD

Root node error: 12/39 = 0.30769

n= 39

        CP nsplit rel error  xerror    xstd
1 0.58333      0   1.00000 1.00000 0.24019
2 0.01000      1   0.41667 0.41667 0.17398
>
```


size of tree
1
2
X-val Relative Error
0.2
0.6
1.0
Inf
0.076
cp


Empresas do MEI
Período: 2020
QTD
>= 1160
< 1160
Ativa
27 / 27
Baixada
7 / 12


Text-Minig - MEI
Período: 2020
ativa
voluntaria
liquidacao
extincao
encerramento
inconsistência
cadastral
Ocorrências
0e+00
5e+05
1e+06
contagem



O que verificamos nos resultados dos testes, é que apesar do ano de 2020 ser completamente atípico, em função das implicações econômicas por causa da Pandemia do Covid-19, o comportamento do ciclo de vida das empresas não foi modificado.

Podemos observar que as empresas do Lucro Real/Presumido são afetadas de maneira significativa como passar dos anos, ao passo que as empresas do Simples Nacional e do MEI, são afetadas de forma mais branda com o passar do tempo.

Contudo, este trabalho deve ser considerado com um ponto de partida para análises mais profundas, tais como:

- Existe vinculação entre o quadro societário das empresas abertas e o das empresas encerradas;
- As empresas do Simples Nacional e do MEI permanecem mais tempo em funcionamento por falta de encerramento;

- Qual é o passivo deixado pelas empresas encerradas;
- Qual o percentual de “Laranjas” envolvidos nas operações de abertura/fechamento das empresas, etc

7. Links

Link para o repositório Github com o conteúdo do trabalho: https://github.com/Itanima/TCC-PUC-MINAS.

Abaixo a descrição dos arquivos do repositório indicado:

- Arvore-TCC - 2020.R: Script para o modelo de Árvore de Classificação – Base de Testes;
- Arvore-TCC.R: Script para o modelo de Árvore de Classificação – Base de Treinamento;
- Graficos - 2020.R: Script para elaboração dos gráficos – Base de Testes;
- Graficos.R: Script para elaboração dos gráficos – Base de Treinamento;
- LR.csv: DataSet das empresas optantes pelo Lucro Real/Presumido;
- LR_MOTIVO.CSV: Tabela dos Motivos de exclusão das empresas optantes pelo Lucro Real/Presumido;
- Library.R: Script para carregamento das bibliotecas necessárias;
- MEI.csv: DataSet das empresas optantes pelo MEI;
- MEI_MOTIVO.CSV: Tabela dos Motivos de exclusão das empresas optantes pelo MEI;
- SN.csv: DataSet das empresas optantes pelo Simples Nacional;
- SN_MOTIVO.CSV: Tabela dos Motivos de exclusão das empresas optantes pelo Simples Nacional;
- TCC-Data.RData: Base de Dados gerada pelo R;
- base_polaridade.csv: Tabela de polaridades para o Modelo de Mineração de Texto;
- main.R: Script para importação e geração dos DataSet’s;
- sw.csv: Tabela de StopWords para o Modelo de Mineração de Texto;

- text mining - 2020.R: Script para o modelo de Mineração de Texto – Base de Testes;
- text mining.R: Script para o modelo de Mineração de Texto – Base de Treinamento;
- vídeo_da_apresentação.mp4: Vídeo com a apresentação;
- Apresentação.ppsx: Apresentação em PowerPoint;
- TCC Ciência de Dados 2020.Doc: Trabalho de conclusão em Word;
- TCC Ciência de Dados 2020.pdf: Trabalho de conclusão em pdf.

REFERÊNCIAS

Anaconda. Disponível em < https://www.anaconda.com/>.

JUPYTER. Disponível em: <https://jupyter.org/>.

PYTHON. Disponível em: <https://www.python.org/>.

SILGE, Julia; ROBINSON, David. Text Mining with R: Tidy Approach. Sebastopol, Ca: O'reilly, 2017. 194 p.

T.M. Therneau, J.A. Elizabeth. An Introduction to Recursive Partitioning Using the RPART Routines. Mayo Foundation, 2019.

The R Project for Statistical Computing. Disponível em: < https://www.r-project.org/

APÊNDICE

Tabelas

Tabelas de Código do motivo x Motivo da baixa do estabelecimento

Empresas optantes pelo Lucro Real/Presumido

COD_ MOTIVO	MOTIVO
1	ATIVA
2	INEXISTENTE DE FATO
3	OMISSAO DE DECLARACOES
4	INTERRUPCAO TEMPORARIA DAS ATIVIDADES
5	PEDIDO DE BAIXA INDEFERIDA
6	EXTINCAO POR ENCERRAMENTO LIQUIDACAO VOLUNTARIA
7	INCORPORACAO
8	EXTINCAO - TRATAMENTO DIFERENCIADO DADO AS ME E EPP (LEI COMPLEMENTAR 123/2006)
9	INAPTIDAO (LEI 11.941/2009 ART.54)
10	OMISSAO CONTUMAZ
11	PRATICA IRREGULAR DE OPERACAO DE COMERCIO EXTERIOR
12	LOCALIZACAO DESCONHECIDA
13	BAIXA INICIADA EM ANALISE
14	BAIXA DEFERIDA PELA RFB
15	INCONSISTÊNCIA CADASTRAL
16	ANULACAO POR MULTICIPLIDADE
17	ANULACAO POR VICIOS
18	REGISTRO CANCELADO
19	ENCERRAMENTO DA FALENCIA
20	BAIXA INDEFERIDA PELA RFB E AGUARDANDO ANALISE DO CONVENENTE
21	BAIXA INDEFERIDA PELA RFB
22	INDICIO INTERPOS. FRAUDULENTA
23	ANULACAO DE INSCRICAO INDEVIDA
24	CISAO TOTAL
25	EXTINCAO POR ENCERRAMENTO LIQUIDACAO
26	ENCERRAMENTO DA LIQUIDACAO
27	BAIXA REGISTRADA NA JUNTA
28	FUSAO
29	DETERMINACAO JUDICIAL
30	EXTINCAO-UNIFICACAO DA FILIAL
31	OMISSA NAO LOCALIZADA
32	BAIXA DEFERIDA PELA RFB E INDEFERIDA PELO CONVENENTE

33	ELEVACAO A MATRIZ
34	INAPTIDAO
35	BAIXA DEFERIDA PELA RFB E SEFAZ
36	BAIXA INDEFERIDA PELA RFB E DEFERIDA PELO CONVENENTE
37	OMISSA CONTUMAZ
38	FALTA DE PLURALIDADE DE SOCIOS
39	BAIXA DE PRODUTOR RURAL
40	BAIXA DEFERIDA PELA RFB E SEFIN E INDEFERIDA PELA SEFAZ
41	BAIXA DEFERIDA PELA RFB E SEFIN
42	BAIXA DEFERIDA PELA RFB E SEFAZ E INDEFERIDA PELA SEFIN
43	BAIXA INDEFERIDA PELA RFB E SEFAZ
44	ANULACAO POR NAO CONFIRMADO ATO DE REGISTRO DO MEI NA JUNTA COMERCIAL
45	BAIXA INDEFERIDA PELA RFB E SEFAZ E DEFERIDA PELA SEFIN

Empresas optantes pelo MEI

COD_ MOTIVO	**MOTIVO**
1	ATIVA
2	OMISSAO DE DECLARACOES
3	BAIXA INICIADA EM ANALISE
4	PEDIDO DE BAIXA INDEFERIDA
5	ANULACAO POR MULTICIPLIDADE
6	INTERRUPCAO TEMPORARIA DAS ATIVIDADES
7	EXTINCAO POR ENCERRAMENTO LIQUIDACAO VOLUNTARIA
8	LOCALIZACAO DESCONHECIDA
9	REGISTRO CANCELADO
10	INCONSISTÊNCIA CADASTRAL
11	INEXISTENTE DE FATO
12	EXTINCAO - TRATAMENTO DIFERENCIADO DADO AS ME E EPP (LEI COMPLEMENTAR 123/2006)
13	BAIXA INDEFERIDA PELA RFB
14	INDICIO INTERPOS. FRAUDULENTA
15	DETERMINACAO JUDICIAL
16	EXTINCAO-UNIFICACAO DA FILIAL
17	BAIXA INDEFERIDA PELA RFB E DEFERIDA PELO CONVENENTE
18	BAIXA DEFERIDA PELA RFB
19	PRATICA IRREGULAR DE OPERACAO DE COMERCIO EXTERIOR
20	ANULACAO POR VICIOS
21	ANULACAO POR NAO CONFIRMADO ATO DE REGISTRO DO MEI NA JUNTA COMERCIAL
22	BAIXA INDEFERIDA PELA RFB E AGUARDANDO ANALISE DO CONVENENTE
23	INVÁLIDO

Empresas optantes pelo Simples Nacional

COD_ MOTIVO	MOTIVO
1	ATIVA
2	INEXISTENTE DE FATO
3	INTERRUPCAO TEMPORARIA DAS ATIVIDADES
4	OMISSAO DE DECLARACOES
5	PEDIDO DE BAIXA INDEFERIDA
6	EXTINCAO POR ENCERRAMENTO LIQUIDACAO VOLUNTARIA
7	LOCALIZACAO DESCONHECIDA
8	BAIXA INICIADA EM ANALISE
9	BAIXA DEFERIDA PELA RFB
10	BAIXA INDEFERIDA PELA RFB E AGUARDANDO ANALISE DO CONVENENTE
11	BAIXA INDEFERIDA PELA RFB E DEFERIDA PELO CONVENENTE
12	EXTINCAO - TRATAMENTO DIFERENCIADO DADO AS ME E EPP (LEI COMPLEMENTAR 123/2006)
13	DETERMINACAO JUDICIAL
14	INCONSISTÊNCIA CADASTRAL
15	INDICIO INTERPOS. FRAUDULENTA
16	PRATICA IRREGULAR DE OPERACAO DE COMERCIO EXTERIOR
17	BAIXA INDEFERIDA PELA RFB
18	BAIXA DEFERIDA PELA RFB E SEFAZ
19	ANULACAO POR VICIOS
20	EXTINCAO-UNIFICACAO DA FILIAL
21	BAIXA REGISTRADA NA JUNTA
22	FALTA DE PLURALIDADE DE SOCIOS
23	INCORPORACAO
24	OMISSA CONTUMAZ
25	OMISSA NAO LOCALIZADA
26	BAIXA DEFERIDA PELA RFB E INDEFERIDA PELO CONVENENTE
27	ANULACAO DE INSCRICAO INDEVIDA
28	ANULACAO POR MULTICIPLIDADE
29	REGISTRO CANCELADO
30	BAIXA DEFERIDA PELA RFB E SEFIN
31	ELEVACAO A MATRIZ
32	ENCERRAMENTO DA LIQUIDACAO