Trabalho Final Curso Especialização Bi-Master

Análise dos Vencedores de Licitações Públicas por Agrupamento

Aluno: Glauco Pires Rabello
Matrícula: 191.477.007
2019.1
Orientador: Felipe Borges
CCE PUC-Rio

Abril/2021

Objetivo:

Estudar o comportamento e padrões do dataset dos vencedores das licitações públicas, apontar similaridade e outliers.

Fontes informacionais

Para este estudo foram usadas as bases de dados contidas nos portais relacionados abaixo. Não foram usadas quaisquer outras bases de dados que não sejam públicas.

Portal da Transparência
http://www.portaltransparencia.gov.br/download-de-dados

Dados públicos de CNPJ

https://www.receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj

Dados Abertos

https://dados.gov.br/

Softwares ou Ferramentas Usados

Em todo o trabalho foram usados sempre softwares open-source ou sem restrições na utilização de qualquer licença.

MySQL Community
Python

Estruturação Informacional

Foram baixados e carregados os seguintes arquivos para construção de uma base de amostras para executar os testes.

Licitações
- ItemLicitação ; lista dos vencedores por licitação
- Licitações ; lista de participantes por licitação

Sanções
- Empresas Inidoneas
- Empresas Impedidas
- Empresas Punidas
- Acordo Leniencia

Receita Federal
- Base completa de CNPJ – referencia nov/2020
- Dominios: Motivo Situação Cadastral, CNAE e qualificação sócio representante

Foi realizado todo um trabalho de carga, tratamento dos campos e informações com as dimensões para transformar os dados e gerar uma saída (tabela/arquivo) para posterior análise.

O período de informações extraídas compreendeu do período jan-out 2020.

Foi criado um arquivo com estes dados tratados: amostra_cnpj_202001.zip

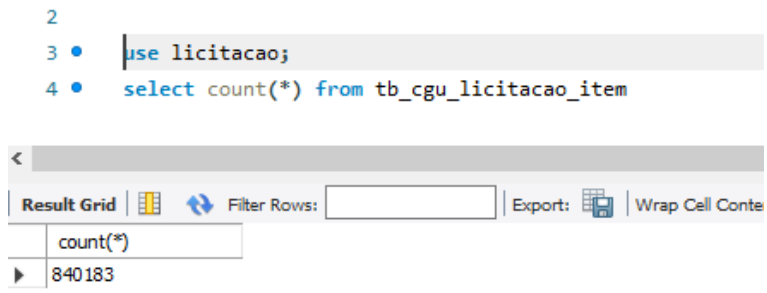

Figura 1: Quantidade de registros na tabela de licitações

Figura 2: Quantidade de registros na tabela de CNPJs

O código da extração final está disponível no arquivo *licitacao_criacao_amostras.sql*

Dicionário Dados da Amostra

Field	Type	Descricao
mesref	char(6)	mes de referencia da licitacao
num_licitacao	varchar(255)	Número que identifica a licitação SIASG
cod_ug	varchar(255)	Número do processo da licitação
dat_resultado	date	data do resultado
modal_compra	varchar(255)	modalidade de compra na licitacao
objeto	text	objeto da licitacao
cnpj	varchar(255)	cnpj do licitante vencedor
tipo_pessoa	varchar(2)	tipo de pessoa NO
ind_matriz	char(1)	indicador matriz ou filial
razao_social	varchar(150)	razao social do licitante
situacao_cadastral	varchar(8)	descricao situacao cadastral
dat_sit_cadastral	date	data situacao cadastral
ano_sit_cadastral	int(4)	ano situacao cadastral
motiv_sit_cadastral	varchar(255)	motivo situacaocadastral se nao ativa
tipo_nat_juridica	varchar(255)	tipo nat juridica
dat_ini_ativ	date	data inicio atividade
ano_ini_ativ	int(4)	ano inicio atividade
qualif_resp	varchar(255)	qualificacao do responsavel da empresa
setor_cnae	varchar(255)	setor cadastro CNAE
porte_empr	varchar(13)	porte empresarial
opt_simples	varchar(11)	optante pelo simples
motiv_impedimento	varchar(255)	motivo de eventual impedimento
motiv_punicao	varchar(255)	motivo punicao
motiv_inidonea	varchar(255)	motivo inidonea

Análise Exploratória

Nesta etapa, já com os dados carregados no banco de dados, foi realizado fase uma análise exploratória para identificar quais dados poderiam ser incluídas nas análises seguintes.

```
PlotBarsColumn('modal_compra', tipo='barh', stack=1, altura=10, fontsize=25 )
```

```
campo   :  modal_compra
--------------------------------------------------------------------------------
amostras :  28219
--------------------------------------------------------------------------------
missing  :  0

- Valores ------------------------------------------------------------------------
Pregão - Registro de Preço      18311
Dispensa de Licitação            4395
Pregão                           2794
Inexigibilidade de Licitação     2613
Tomada de Preços                   72
Convite                            22
Concorrência                       12
Name: modal_compra, dtype: int64

- Ref Cruzada --------------------------------------------------------------------
situacao_cadastral              Ativa   Baixada    Inapta  Suspensa
modal_compra
Concorrência                 1.000000  0.000000  0.000000  0.000000
Convite                      1.000000  0.000000  0.000000  0.000000
Dispensa de Licitação        0.997497  0.002048  0.000000  0.000455
Inexigibilidade de Licitação 0.997704  0.002296  0.000000  0.000000
Pregão                       0.996063  0.003221  0.000000  0.000716
Pregão - Registro de Preço   0.998853  0.000765  0.000273  0.000109
Tomada de Preços             1.000000  0.000000  0.000000  0.000000
```

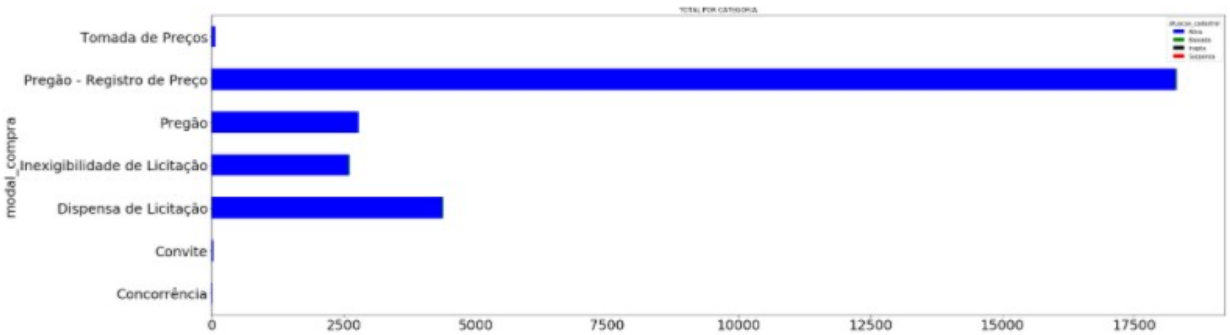

Figura 3: Exemplo passo da análise exploratória do campo modalidade de compra

Códigos contidos no arquivo 1-*licitacoes_cnpj_analise_exploratoria.ipynb*

Com os dados disponíveis foi iniciada a modelagem para a identificação de possíveis similaridades.

Out[93]:

	mesref	num_licitacao	cod_ug	dat_resultado	modal_compra	objeto	cnpj	tipo_pessoa	ind_matriz	razao_social	...	dat_sit_cadastral	ano
0	202001	000012018	925206	2020-01-21	Pregão	Pregão Eletrônico - Contratação de empresa es...	55905350000199	PJ	1	PAINEIRAS LIMPEZA E SERVICOS GERAIS LTDA	...	2005-11-03	
1	202001	000012018	160012	2020-01-22	Pregão - Registro de Preço	Pregão Eletrônico - Aquisição de alimentos pa...	21860768000105	PJ	1	W SANTOS CHAVES	...	2015-02-11	
2	202001	000012018	160012	2020-01-22	Pregão - Registro de Preço	Pregão Eletrônico - Aquisição de alimentos pa...	30771627000107	PJ	1	E DA SILVA PINTO COMERCIO	...	2018-06-23	
3	202001	000012018	160012	2020-01-22	Pregão - Registro de Preço	Pregão Eletrônico - Aquisição de alimentos pa...	28388146000175	PJ	1	ANDREA DA COSTA FERREIRA EIRELI	...	2017-08-10	
4	202001	000012018	160012	2020-01-22	Pregão - Registro de Preço	Pregão Eletrônico - Aquisição de alimentos pa...	28388146000175	PJ	1	ANDREA DA COSTA FERREIRA EIRELI	...	2017-08-10	

Figura 4: Amostra dos dados disponíveis no Jupyter

Análise variável ‘Modalidade de Compras’

Como esta variável é a mais importantes do dataset, pois é a que define o motivo da licitação, foi realizado um teste no dataset mantendo esta informação para confirmação.

No processo de construção de modelo, para melhor a analise convertendo as váriaveis categóricas em numéricas por Encoding e depois aplicando o método de PCA.

Nesta etapa descobrimos algumas variáveis que deverão definir este modelo. (foi escolhido 80% para explicar esses dados)

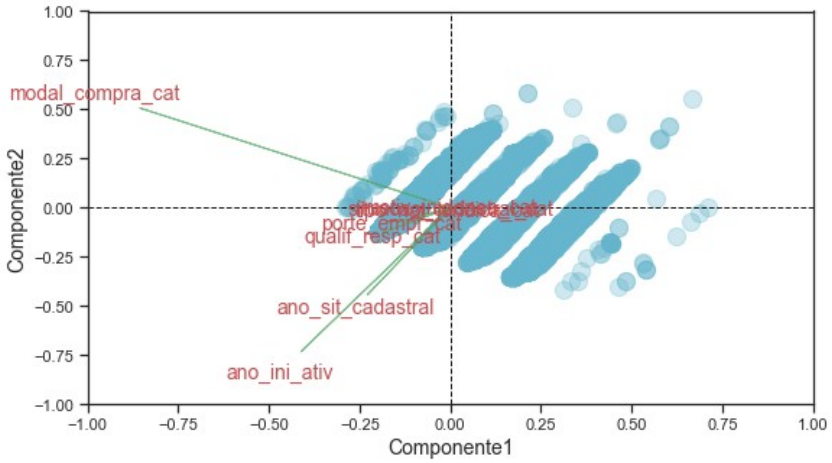

Figura 5: Variancias dos dados do dataset após aplicação do PCA

Teste com método K-Means

Este teste foi gerado para identificar se seria possível usá-lo, pois é um método de fácil implantação. Desta maneira verificamos o K através da inércia do modelo e analisando se o modelo gerado trará bons resultados.

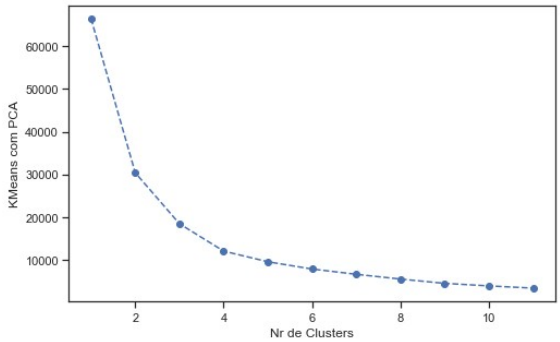

Figura 6: Gráfico de inércia por quantidade de clusters K

Após a execução do K-Means, visualmente chegamos a este resultado que pode ser facilmente interpretado como um método ruim para este dataset.

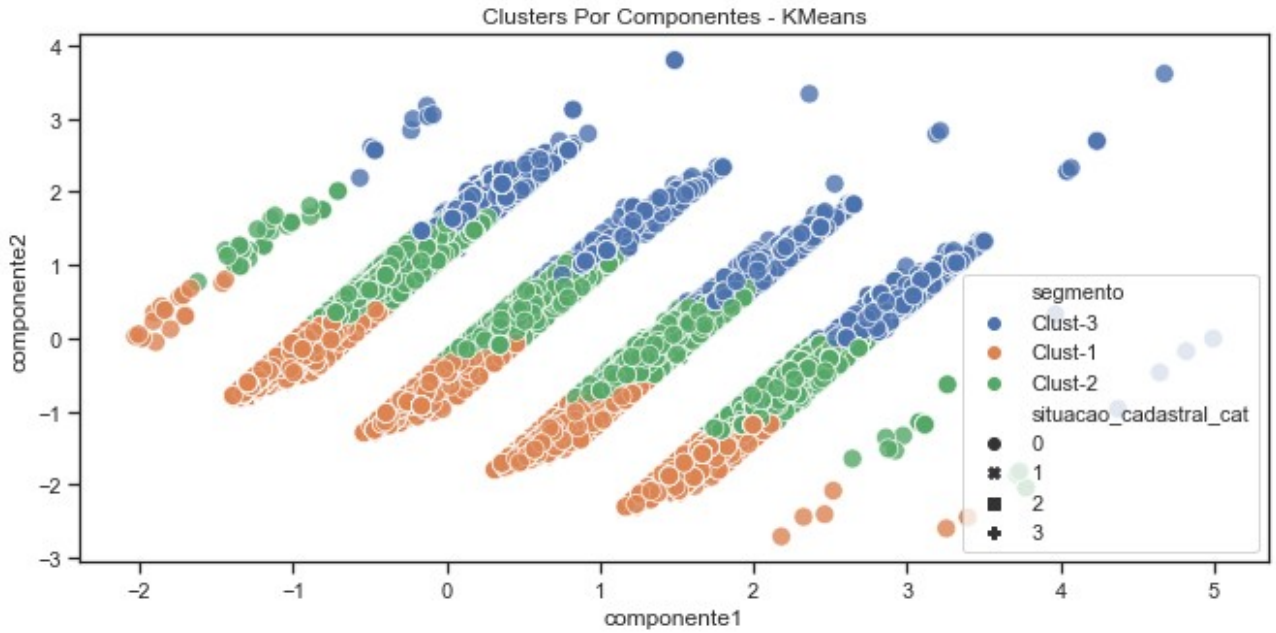

Figura 7: Resultado do algoritmo K-Means

Algoritmo DBSCAN

Escolha deste método por não ter uma quantidade de clusters pré-definidos e encontrar quase toda as formas.

Através de algumas inferências podemos escolher as parametrizações melhores.

Chegamos a um ótimo resultado desde já.

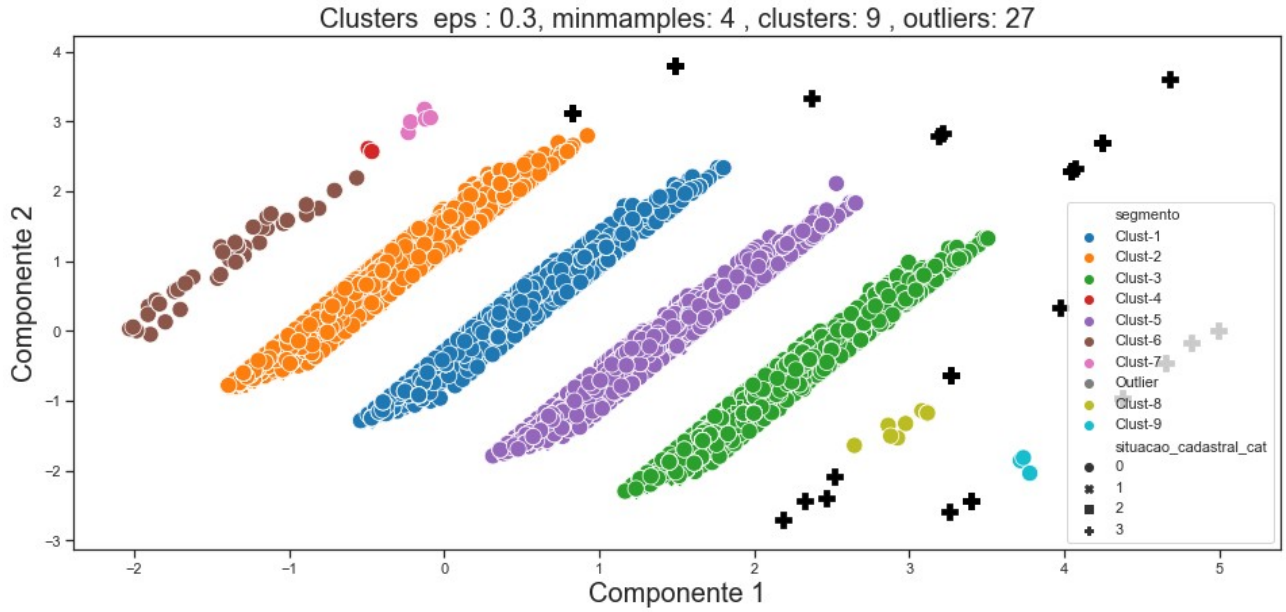

Figura 8: Resultado do algoritmo DBSCAN

Uma reparametrização no algoritmo pode melhorar este resultado para um ajuste fino. Observando que é necessário equilibrar a quantidade de clusters e outliers e as frequências individuais dos clusters.

```
# Quantidade por Cluster Gerados
modelo.segmento.value_counts()
```

```
Clust-2    18307
Clust-3     4389
Clust-1     2793
Clust-5     2611
Clust-6       62
Outlier       27
Clust-8       15
Clust-7        6
Clust-9        5
Clust-4        4
```

Figura 9: Totais por cluster

Alguns outliers gerados foram identificados desta maneira, pois estavam contidos em uma faixa com poucas amostras. Estes outliers também poderiam ser identificados como um pequeno cluster rotulado.

Conforme figura abaixo, podemos questionar a parametrização do modelo e realizar ajustes.

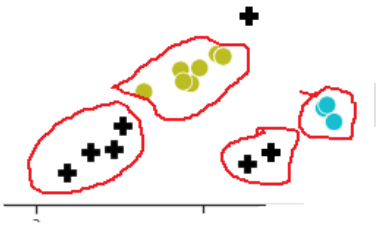

Figura 10: Faixa de outliers e clusters com poucas amostras

A variável **ano_ini_atividade** tambem ajudou na identificação de alguns outliers, mas pontos mais isolados.

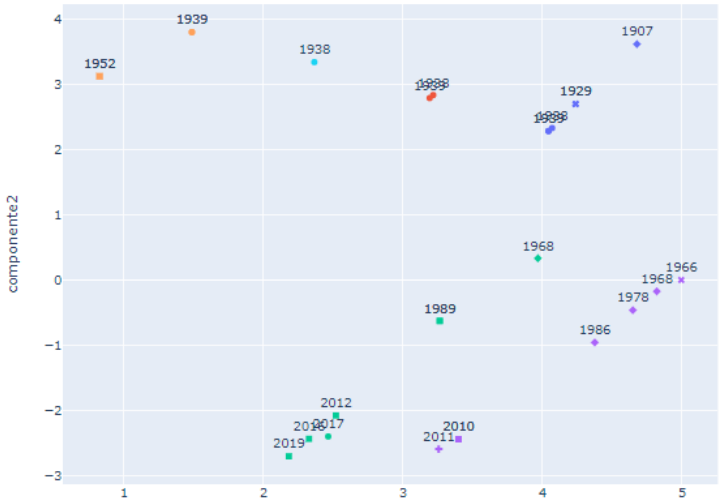

Figura 11: Outilers

Assim concluímos que os clusters identificados (Gráfico 8) se encaixam “exatamente” com as amostras de Modalidade de Compras.

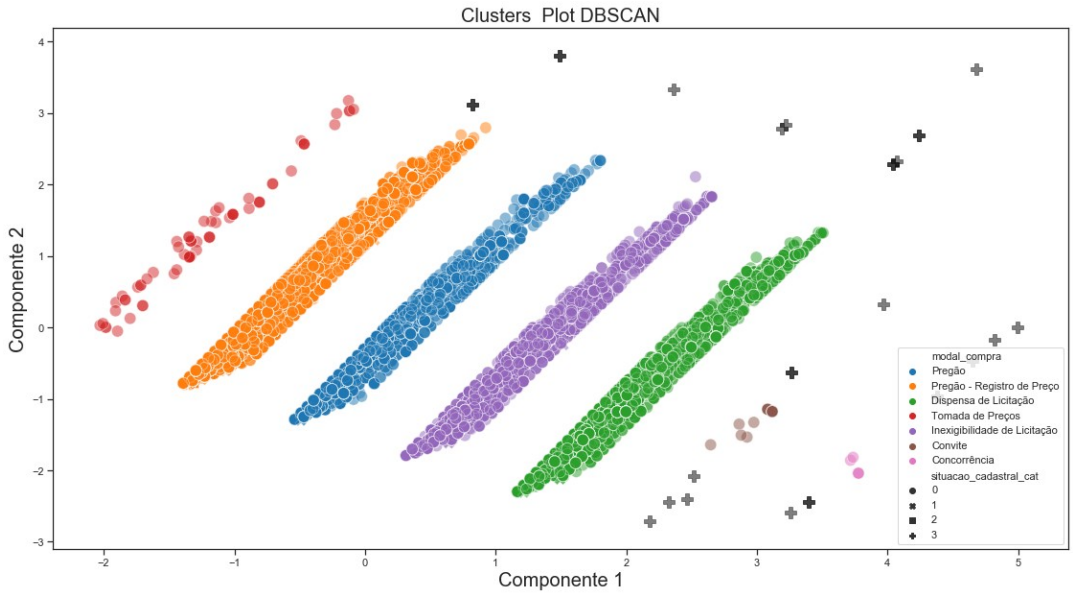

Figura 12: Dados identificados por Modalidade de Compras

Códigos disponíveis no arquivo 2-*licitacoes_cnpj_agrupamento_todas_modalidades.ipynb*

Analise por Modalidade de Compra

Foi escolhido a modalidade com maior quantidade de amostras para prosseguimento das analise.

```
df.modal_compra.value_counts()
```

Pregão - Registro de Preço	18311
Dispensa de Licitação	4395
Pregão	2794
Inexigibilidade de Licitação	2613
Tomada de Preços	72
Convite	22
Concorrência	12

Figura 13: Quantidade de amostras por Modalidade de Compra

Os códigos desta analise estão no arquivo *licitacoes_cnpj_agrupamento_por_modal_compra.ipynb*

Após a execução do algoritmo DBSCAN foi encontrado este resultado.

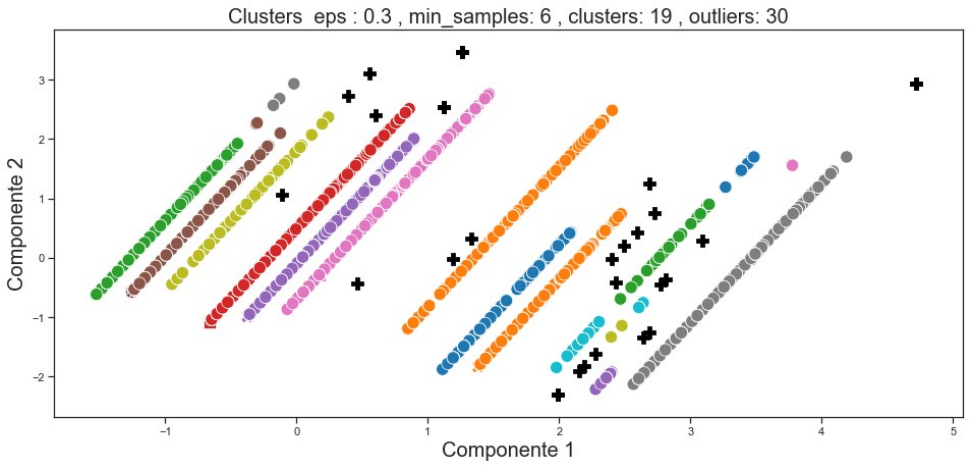

Figura 14: Resultado do DBSCAN por 'Pregão - Registro de Preço'

Foram necessárias várias re-execuções deste algorimo, após a identificação de valores de eps e min amostras, reparametrizando-os, pois foi identificado uma possível melhoraria dos resultados (menos clusters, menos outliers)

Inicialmente usamos um valor de eps muito baixo cujo resultado ficou bem ruim, com muitos clusters.

Análise Resultados

Após a execução esse resultado foi gerado

```
# Quantidade por Cluster Gerados
modelo.segmento.value_counts()
```

```
Clust-4     4630
Clust-3     3696
Clust-5     2658
Clust-6     2208
Clust-2     1269
Clust-7     1168
Clust-1      982
Clust-8      506
Clust-9      471
Clust-12     354
Clust-13     153
Clust-10      73
Clust-11      38
Outlier       30
Clust-14      19
Clust-16      16
Clust-19      11
Clust-15      11
Clust-17      10
Clust-18       8
```

Figura 15: Totais de clusters gerados

Estas duas variáveis foram identificadas como principais, boa distribuição de frequências.

```
modelo.qualif_resp.value_counts( normalize=True )
```

```
Sócio-Administrador                0.462072
Titular Pessoa Física Residente    0.349407
Empresário                         0.123205
Administrador                      0.030364
Presidente                         0.019333
Diretor                            0.014418
Administrador Judicial             0.000928
Inventariante                      0.000109
Tabelião                           0.000055
Sócio-Gerente                      0.000055
Síndico (Condomínio)               0.000055
Procurador                         0.000000
Fundador                           0.000000
Name: qualif_resp, dtype: float64
```

```
modelo.porte_empr.value_counts(normalize=True )
```

```
Pequeno Porte    0.511277
Micro Empresa    0.336410
Demais           0.152313
Name: porte_empr, dtype: float64
```

Figura 16: Frequências por Qualif_resp e porte_empre

Foi criado um gráfico em função destas duas variáveis.

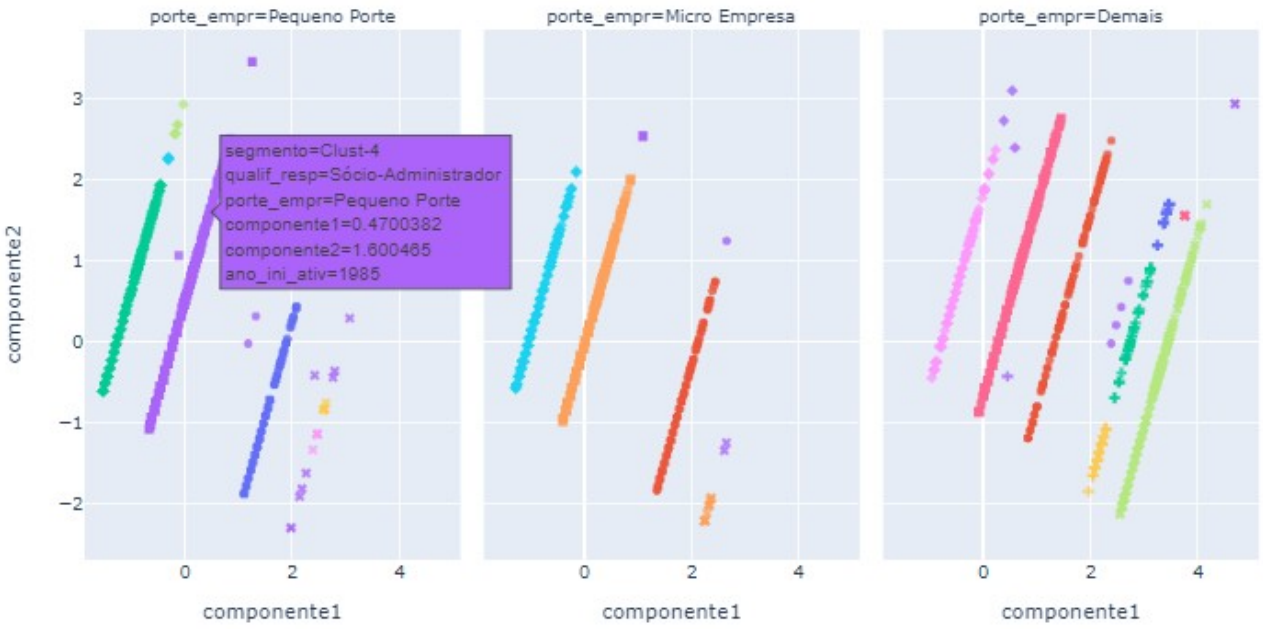

Figura 17: Gráfico resultante por qualificação responsável e porte empresa

No gráfico acima, ver no Jupyter, que os clusters ficaram bem definidos com estas duas variáveis. Como ainda pode-se identificar na tabela abaixo que as segmentações foram criadas.

:

	segmento	Clust-1	Clust-10	Clust-11	Clust-12	Clust-13	Clust-14	Clust-15	Clust-16	Clust-17	Clust-18	Clust-19	Clust-2	Clust-3	Clust-4	Clust-5	Clust-6	Clust-7	Cl
porte_empr	qualif_resp																		
Demais	Administrador	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	Administrador Judicial	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	
	Diretor	0	73	38	0	153	0	0	0	0	0	0	0	0	0	0	0	0	
	Empresário	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	Presidente	0	0	0	354	0	0	0	0	0	0	0	0	0	0	0	0	0	
	Síndico (Condomínio)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	Sócio-Administrador	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1168	
	Tabelião	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	Titular Pessoa Física Residente	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Micro Empresa	Administrador	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	
	Empresário	0	0	0	0	0	0	0	0	0	0	0	1269	0	0	0	0	0	
	Sócio-Administrador	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2658	0	0	
	Titular Pessoa Física Residente	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2208	0	
Pequeno Porte	Administrador	0	0	0	0	0	0	0	0	0	8	11	0	0	0	0	0	0	
	Administrador Judicial	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	Empresário	982	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	Sócio-Administrador	0	0	0	0	0	0	0	0	0	0	0	0	0	4630	0	0	0	
	Titular Pessoa Física Residente	0	0	0	0	0	0	11	0	10	0	0	0	3696	0	0	0	0	
	Inventariante	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	Sócio-Gerente	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figura 18: Tabela dados resultados por qualificação responsável e porte empresa

Outliers

Da mesma maneira da análise com dataset completo, os clusters gerados tem a mesma caracteristica como identificamos visualmente

- A variável ‘ano inicio atividade’ foi um dos fatores de separação;
- Alguns clusters tem a mesma quantidade de amostras de pequenos outliers abaixo, será necessário algum ajuste fino no algoritmo para incluir alguns destes registros.

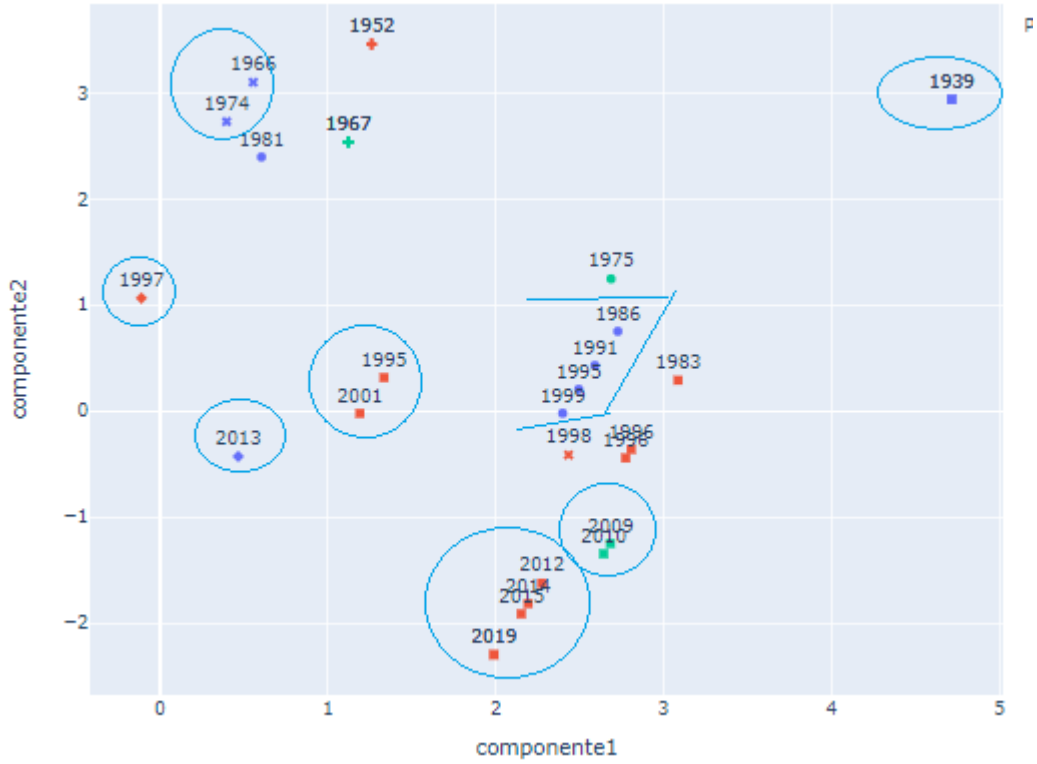

Figura 19: Gráfico de outliers

Conclusões finais

Nesta etapa foram identificados clusters por Modalidades de Compra.

Caso selecionamos outra modalidade (ou adicionar outros filtros) o resultado gerado pode ser diferente. Nestes casos sempre serão necessários a reparametrização diferente no DBSCAN.

Conseguimos demonstrar que a clusterização é viável com as variáveis disponíveis neste dataset publico e realizar a segmentação com as similaridades descobertas.

Sempre que houver alguma necessidade adicional o algortimo pode ser reexecutado diversas vezes , com novos parametros para melhorar mais o resultado. Haverá um custo de processamento maior se a quantidade de amostras do dataset for maior e na quantidade de iterações necessárias para comparação dos resultados.

Ainda foi percebido que a análise pode ter continuidade, adicionando mais variáveis e um dataset maior para que possa melhorar a qualidade dos resultados e reduzindo outliers. Ainda posteriormente pode-se adicionar outros métodos supervisionados de análise para previsão com novas entradas de dados.

-x-