Nanodegree Engenheiro de Machine Learning

Projeto final

Alessandre Martins

31 de dezembro de 2050

Sumário

I. Definição

Neste projeto iremos analisar as empresas de saneamento quanto ao relacionamento entre a estrutura de propriedade e a eficiência operacional obtida por diversos atributos.

Visão geral do projeto

Para compor o ranking de empresas, consideramos os conjunto de dados fornecidos pelo Sistema Nacional de Informações sobre Saneamento (SNIS)[1] e pelo Instituto Trata Brasil [2]. Iremos selecionar os 100 municípios mais populosos do Brasil em 2016, para explorar um pequeno subconjunto de dados como amostra e identificar se alguma empresa está altamente correlacionada com outra. Em seguida vamos pré-processar os dados, dimensionando cada Natureza Jurídica e remover os outliers caso exista algum. Iremos aplicar o PCA(Principal Component Analysis[3]) e um algoritmo de clustering para criar os segmentos das empresas selecionadas no conjunto de dados limpo anteriormente. Também iremos estimar a efetividade operacional de cada empresa, conforme o método DEA(Data Envelopment Analysis[4]) e analisar a relação entre os resultados do DEA e a estrutura de propriedade[5]. As empresas de cada município brasileiro do setor de saneamento básico foram selecionadas, cujos dados referentes foram coletados das demonstrações contábeis. Como resultado deste projeto, procura-se enfatizar os problemas encontrados mais significativos e analisar as estruturas de propriedades das empresas privadas e estatais do setor de saneamento básico do Brasil. Por fim, vamos comparar os segmentos encontrados com uma marcação adicional e considerar a maneira como essa informação poderia auxiliar os prefeitos dos municípios com futuras tomadas de decisões de serviço de saneamento básico de suas cidades.

A estrutura de propriedade de saneamento básico influencia diretamente o desenvolvimento de uma cidade. Pois o saneamento básico adequado nos dará indicativos do impacto de um município no meio ambiente, se tem condições de prevenir doenças e melhorar a saúde, a qualidade de vida da população. O saneamento básico também pode definir à produtividade do indivíduo elevando a atividade econômica do município.

---

[1] "Sistema Nacional de Informações sobre Saneamento - SNIS." http://app3.cidades.gov.br/serieHistorica/. Acessado em 23 jun. 2018.
[2] "Ranking do Saneamento - As 100 maiores cidades do ... - Trata Brasil." http://tratabrasil.org.br/datafiles/estudos/ranking/2016/tabela-das-100-cidades.pdf. Acessado em 23 jun. 2018.
[3] "Segmentation and Clustering | Udacity." https://br.udacity.com/course/segmentation-and-clustering--ud981. Acessado em 31 mai. 2018.
[4] "Tutorial » Data Envelopment Analysis: DEAzone.com." http://deazone.com/en/resources/tutorial. Acessado em 31 mai. 2018.
[5] "Governança Corporativa e Estrutura de Propriedade no Brasil ...." http://www.pablo.prof.ufu.br/artigos/ebf3.pdf. Acessado em 31 mai. 2018.

Descrição do problema

Diante dos riscos que o setor de saneamento expressa atualmente, qual a correlação entre a estrutura de propriedade das empresas de saneamento básico das principais cidades brasileiras com os níveis de eficiência através de machine learning?

Por causa de problemas de saneamento básico, os municípios brasileiros coordenaram processos de privatização do setor[6] de saneamento, a pretexto de inscrever maior eficiência ao setor. Entretanto, o processo de privatização pode repetir ao serviço público problemas e riscos típicos dos mercados financeiros, com destaque aos conflitos de agência derivados das diferentes estruturas de propriedade[7]. As empresas privadas expressam diferentes níveis de concentração acionária, com possíveis prejuízos à transparência e percepção a respeito da idoneidade dos proprietários, além de fatores relacionados à instabilidade dos mercados e falhas regulatórias. No final das contas, a população usuária do serviço experimenta as consequências de um processo inadequado de privatização, com perda à divulgada eficiência típica do setor privado[8].

Métricas

A estrutura de dados não é clara, caso exista. Portanto, para selecionarmos o melhor método de clusterização iremos quantificar a "eficiência" de um clustering ao calcular o coeficiente de silhueta[9] de cada ponto de dados, fornecendo um método de pontuação simples de um dado clustering. Uma silhueta define a área de contorno de cada cluster, nos mostrando que pontos se localizam dentro do cluster e quais pontos ficam em uma localização no meio de dois clusters.

Também iremos criar uma feature para auxiliar no processo de clusterização calculando o índice de eficiência DEA pela relação entre os insumos e produtos. Quando maior a razão entre os produtos gerados e os insumos empregados, mais eficiente será a DMU, ou seja:

eficiência DEA = y1u1 + y2u2x1v1+ x2v2

Onde u e v são os pesos atribuídos para y e x que são as saídas e insumos respectivamente. Atribuir pesos é algo crítico que influenciam completamente a otimização

---

[6] "Do ownership and size affect the performance of water ... - Springer Link." 2 abr. 2011, https://link.springer.com/article/10.1007/s10997-011-9173-6. Acessado em 23 jun. 2018.

[7] "The Structure of Corporate Ownership: Causes and ... - Jstor." https://www.jstor.org/stable/1833178. Acessado em 23 jun. 2018.

[8] "Privatization and Regulation - unpan1.un.org, 24.07.2012." http://unpan1.un.org/intradoc/groups/public/documents/un/unpan000152.pdf. Acessado em 23 jun. 2018.

[9] (n.d.). sklearn.metrics.silhouette_score — scikit-learn 0.19.1 documentation. Recuperado em junho 23, 2018, pelo URL http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

do problema, o DEA tem mecanismos para contornar isso, eles dizem quais variáveis foram importantes para determinar a eficiência ou ineficiência de cada DMU.

II. Análise

A proposta é analisar as empresas de saneamento das principais cidades brasileiras quanto ao relacionamento entre a estrutura de propriedade e a eficiência operacional usando métodos de aprendizagem não-supervisionada de machine learning. Com base nessa análise, determinar como as empresas de saneamento privada ou pública influenciam na eficiência do saneamento de um município. Pode ser que uma natureza jurídica das empresas se destaque mais que a outra. Também se propõe investigar a estrutura de propriedade das empresas de saneamento básico presentes nas principais cidades do Brasil, estimar a efetividade operacional de cada empresa, conforme o método DEA e analisar a relação entre os resultados do DEA e a estrutura de propriedade.

Exploração dos dados

Exploramos os dados através de visualizações e códigos para entender como cada atributo é relacionado a outros. Observamos as descrições estatísticas do conjunto de dados, considerando a relevância de cada atributo, e selecionando alguns exemplos de pontos de dados do conjunto de dados.

Descrição dos atributos do Dataset:

- Cód Município: Código do Município;
- Tp Empresa: Tipo Empresa;
- Nat Jurídica: Natureza Jurídica;
- Pop Total: População total do município do ano de referência(Habitantes);
- Perdas Fat: Índice de perdas faturamento(%)';
- Perdas Dist: Índice de perdas na distribuição (%);
- Tarifa méd: Tarifa média praticada (Reais/m³);
- Ext rede água: Extensão da rede de água (km);
- Ext rede esgotos: Extensão da rede de esgotos (km);
- Empregados: Quantidade total de empregados próprios
- Município: Cidade;
- UF: Estado;
- Empresa: Operador;
- At Água: Indicador de atendimento urbano de água(%)
- At Esgoto: Indicador de atendimento urbano de esgoto(%)
- Novas Lig Água: Indicador novas ligações de água/ligações faltantes (%);
- Novas Lig Esgoto: Indicador novas ligações de esgoto/ligações faltantes (%);
- Ev Perdas Fat: Indicador evolução nas perdas de faturamento (%)';

- Ev Perdas Dist: Indicador evolução nas perdas de distribuição (%);
- Esg Tratado: Indicador de esgoto tratado por água consumida (%);
- Investimento: Investimento 5 anos (Milhões de Reais)

Para fins estatísticos vamos considerar dois tipos de empresas. As empresas de natureza jurídica 'Privada' são empresas privadas e a empresas de natureza jurídica diferente de 'Privada' serão consideradas como 'Públicas', até mesmo as empresas de economia mista.

Existem 95 pontos no conjunto de dados de 2016 que representam as 95 cidades mais populosas do Brasil. Foram excluídas 5 por não terem informações nos atributos na origem de dados referente a NSIS. A média geral da Eficiência das empresas foi de 99.36%. A mediana das eficiências das empresas foi de 92.39%. 25% das empresas foram abaixo de 71.46% de eficiência. 25% das empresas atingiram 100.00% de eficiência.

Tab1. *Estatísticas descritivas do conjunto de dados.* Resume a tendência central, a dispersão e a forma da distribuição de um conjunto de dados.

	At Esgoto	At Água	Eficiencia_DEA	Empregados	Esg Tratado	Ev Perdas Dist	Ev Perdas Fat	Ext rede esgotos	Ext rede água	Investimento	Novas Lig Esgoto	Novas Lig Água	Perdas Dist	Perdas Fat	Tarifa méd
count	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00
mean	68.20	93.29	99.36	3325.21	51.10	4.07	14.16	1245.59	2047.80	338.04	26.86	53.70	41.84	35.67	3.57
std	29.36	13.31	106.90	2311.07	31.67	9.35	54.28	2019.92	2614.55	954.98	32.70	39.39	12.84	18.39	0.97
min	0.75	30.10	55.67	136.00	0.00	0.00	0.00	12.22	361.00	0.00	0.00	0.00	13.05	-13.14	1.69
25%	45.44	91.35	71.46	1370.00	24.40	0.00	0.00	420.99	915.40	78.72	3.50	12.93	33.49	23.85	2.95
50%	75.29	99.70	92.39	2820.00	51.55	0.00	0.00	687.70	1378.61	159.93	12.92	54.93	40.29	35.61	3.54
75%	94.90	100.00	100.00	4395.00	77.34	4.52	5.57	1344.28	1957.11	290.48	35.03	99.38	49.02	49.06	4.02
max	100.00	100.00	1085.51	8980.00	100.00	69.35	430.86	17101.64	21262.00	9113.98	100.00	100.00	70.88	69.77	6.69

Tab2. *Ranking de Natureza Jurídica/Eficiência.*

Nat Jurídica	Eficiencia_DEA	At Água	At Esgoto	Esg Tratado	Novas Lig Água	Novas Lig Esgoto	Perdas Fat	Perdas Dist	Ev Perdas Fat	Ev Perdas Dist
Autarquia	153.807500	97.050000	85.766875	49.350625	61.198125	42.623750	41.961875	42.008750	5.860000	6.262500
SEMAPriv	92.215000	100.000000	95.875000	87.746667	99.405000	84.825000	24.695000	35.030000	0.891667	0.508333
SEMAPubl	89.155625	91.142187	60.247031	47.195000	47.423750	17.721875	36.138750	43.803281	18.348125	2.995938
Privada	80.003750	97.575000	72.628750	63.870000	57.391250	23.845000	28.376250	32.250000	8.938750	11.208750
Pública	78.660000	96.100000	95.010000	6.880000	31.420000	35.440000	29.640000	31.520000	0.000000	1.800000

Tab3. *Média e mediana da eficiência das empresas de saneamento.*

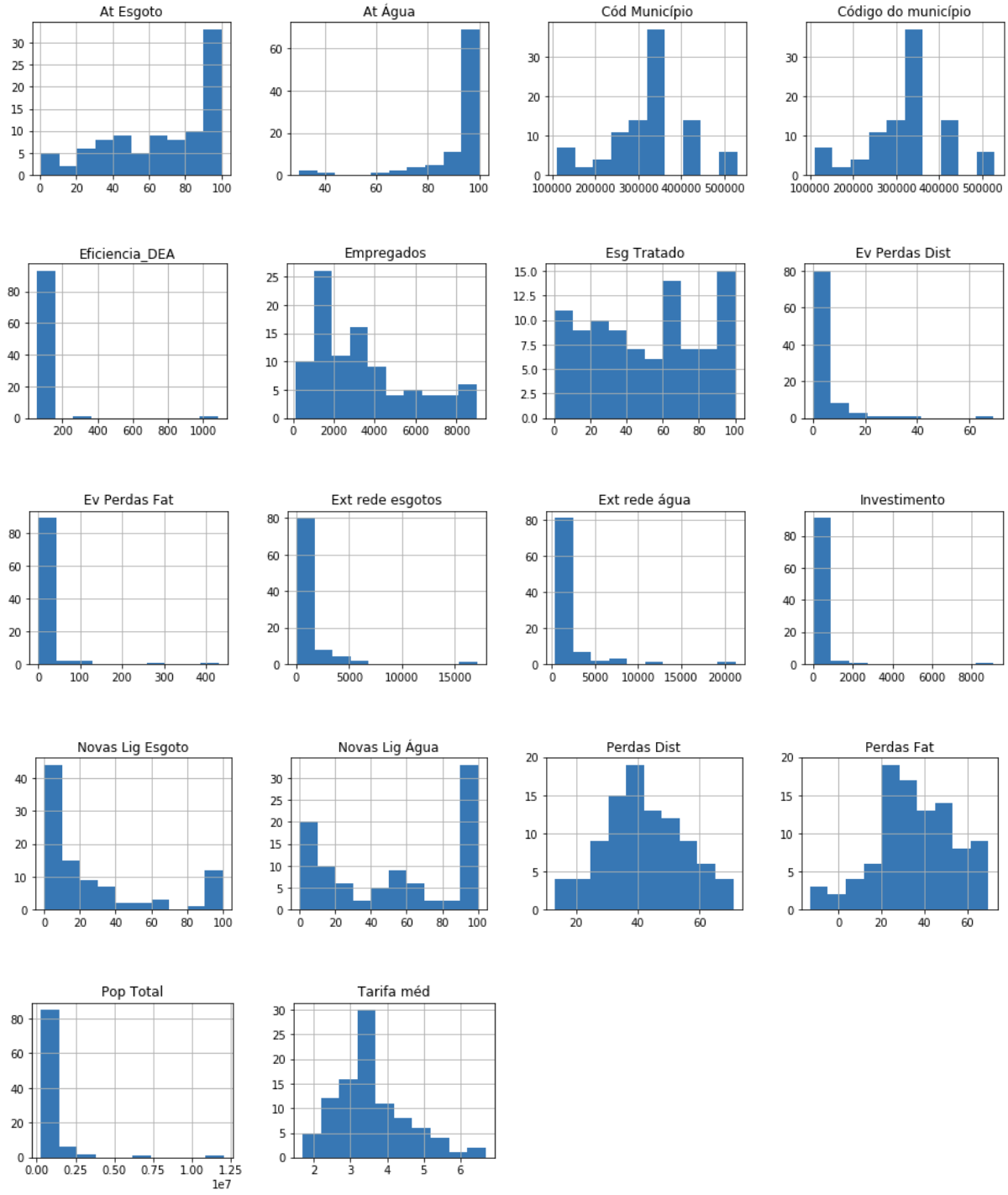
Nat Jurídica	MEDIA	MEDIANA
Autarquia	153.807500	100.000
Privada	80.003750	79.485
Pública	78.660000	78.660
SEMAPriv	92.215000	95.300
SEMAPubl	89.155625	89.290

Tab4. *Amostras selecionadas para acompanhamento do projeto.*

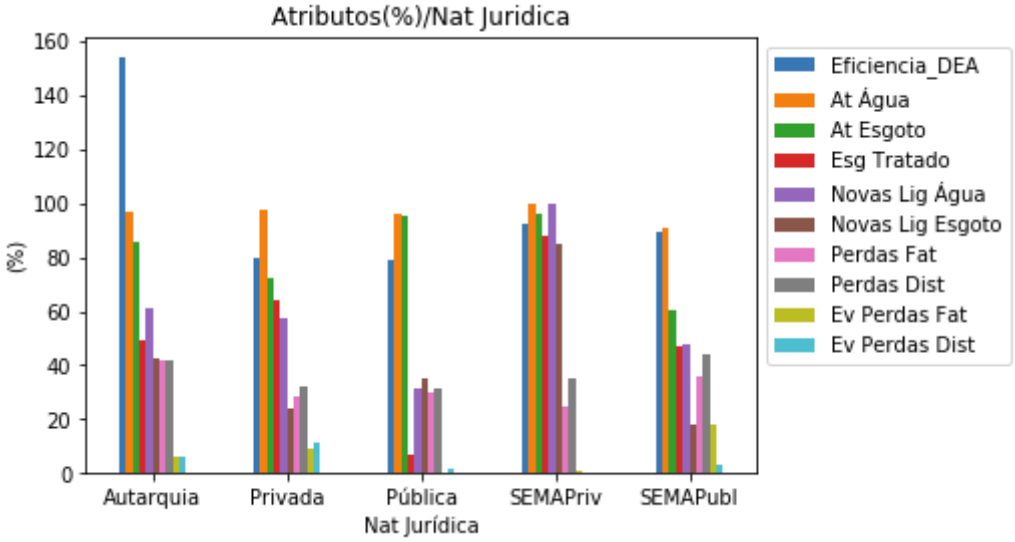
	Perdas Fat	Perdas Dist	Tarifa méd	At Água	At Esgoto	Novas Lig Água	Novas Lig Esgoto	Ev Perdas Fat	Ev Perdas Dist	Esg Tratado	Eficiencia_DEA	Investimento	Ext rede água	Ext rede esgotos	Empregados
0	24.69	36.69	3.66	100.0	97.00	77.37	57.45	0.00	0.00	61.96	65.39	9113.98	21262.0	17101.64	4897
1	52.41	25.36	4.16	99.0	85.16	58.78	17.29	0.84	3.90	44.51	66.43	1922.68	10891.2	4864.51	3105
2	57.09	59.22	3.79	100.0	52.26	52.12	0.48	0.00	7.02	30.90	63.39	582.67	2523.0	905.00	6840

Visualização exploratória

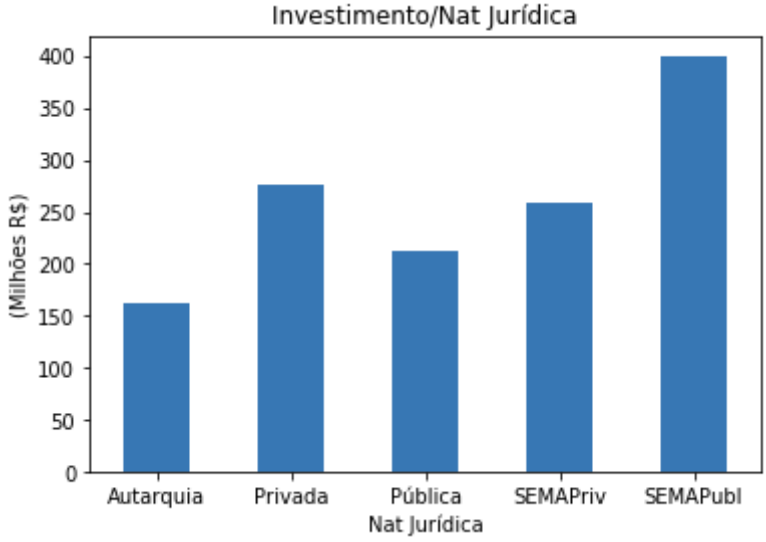
Graf1. *Representação da distribuição de dados.*



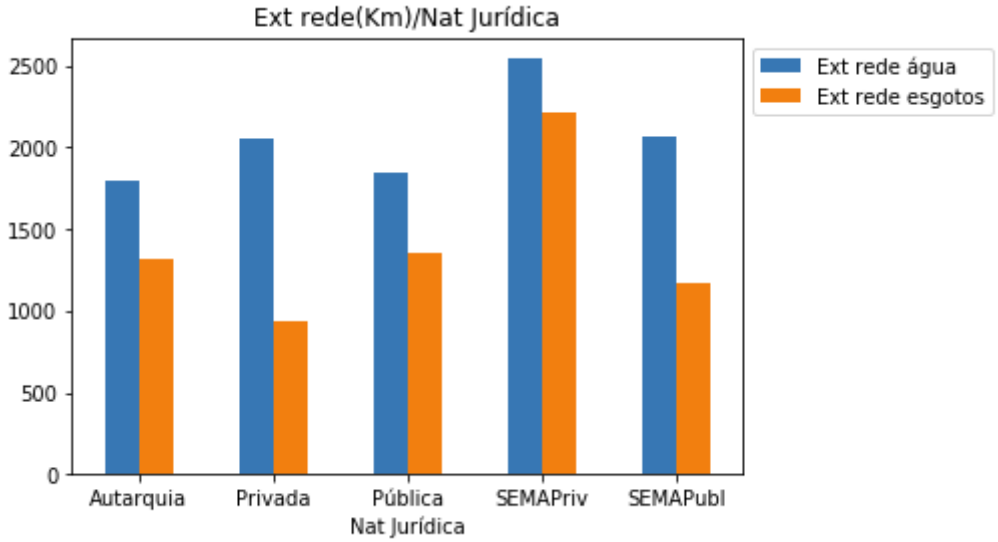
Graf2. *Gráfico que compõe os índices pela Natureza.* Podesse observar que o Índice de Eficiência para Autarquia está acima de 100%, uma discrepância alta que pode enviesar a análise.



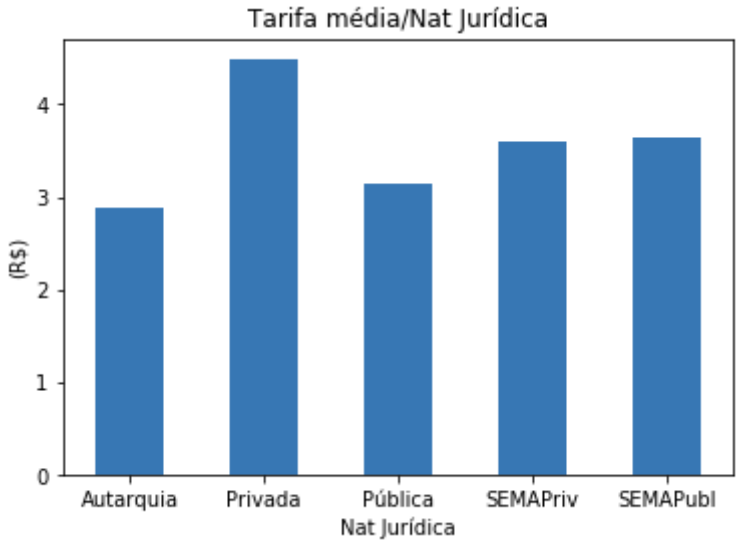
Graf3. *Gráfico da Média de investimento por Natureza.* As empresas de saneamento de Sociedade Econômica Mista de Administração Pública são as que mais recebem investimento.



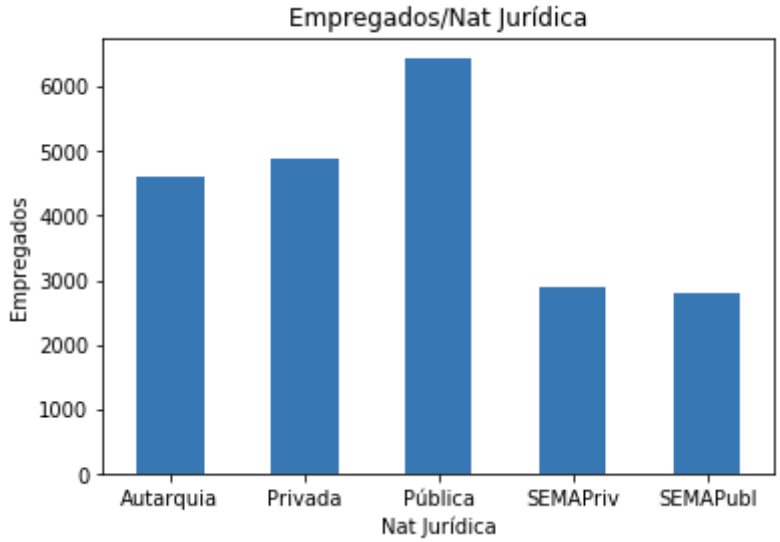
Graf4. *Gráfico com a média das Ext rede de água e esgoto por Natureza.* A empresas que possuem a maior extensão de rede de água e esgoto são as de Sociedade Econômica Mista de Administração Privada. Embora sejam dois atributos correlacionados podemos ver que não tem um crescimento proporcional para as extensões de rede entre as Naturezas. Exemplo disso é a comparação da SEMAPubl com a Pública. A SEMAPubl possui uma extensão de rede de água maior que a Pública, mas a extensão de rede de esgoto é menor que da publica.



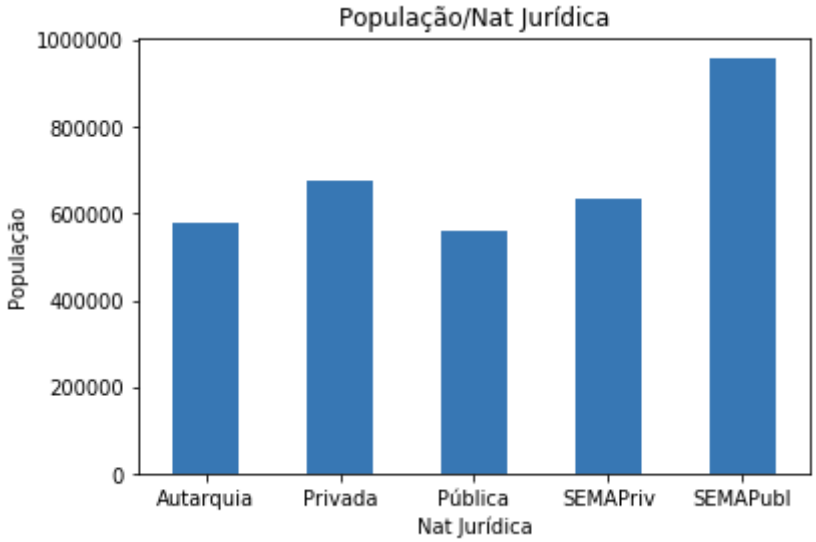
Graf5. *Gráfico da Tarifa média por Natureza.* Apesar das empresas de saneamento privadas receberem menos investimento que as SEMAPubl, podemos detectar a tarifa média da privada é alta comparada com a SEMAPubl.



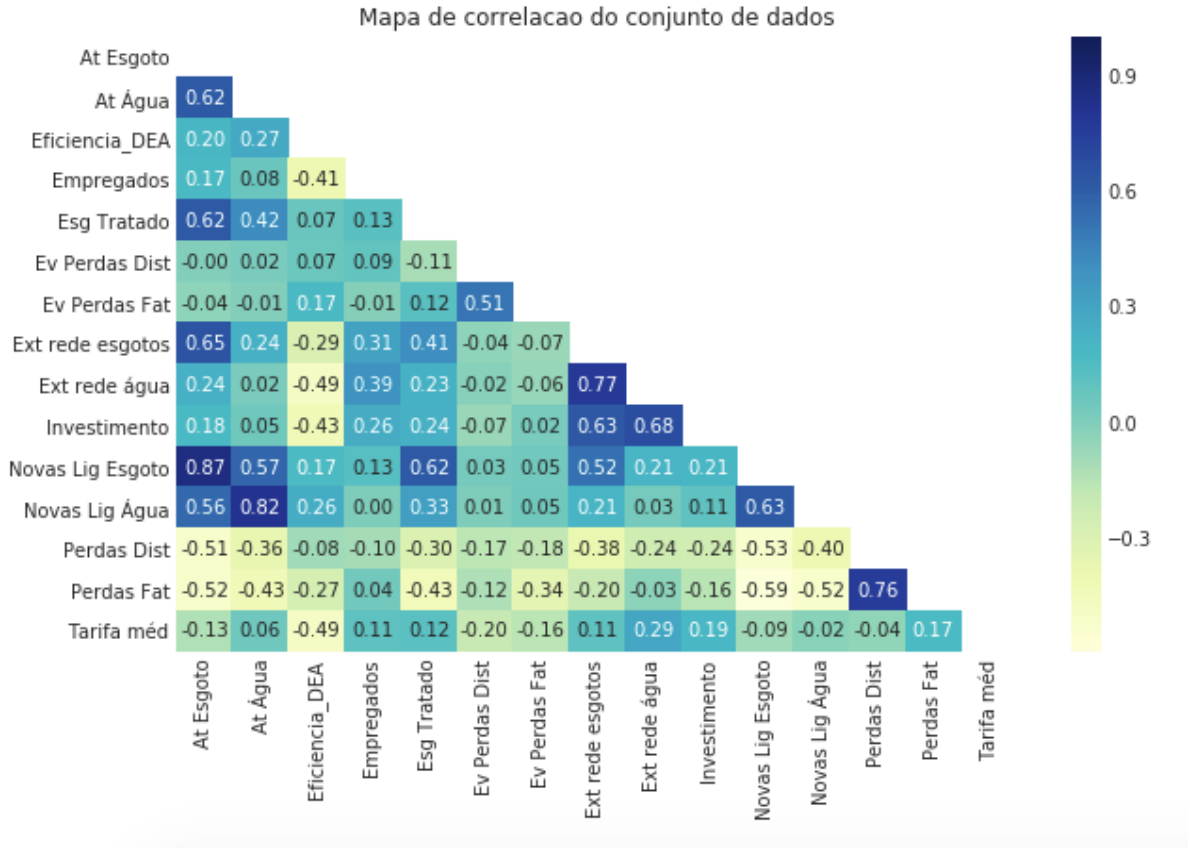
Graf6. *Gráfico Empregados por Natureza.* As empresas públicas estão no ranking de funcionários.



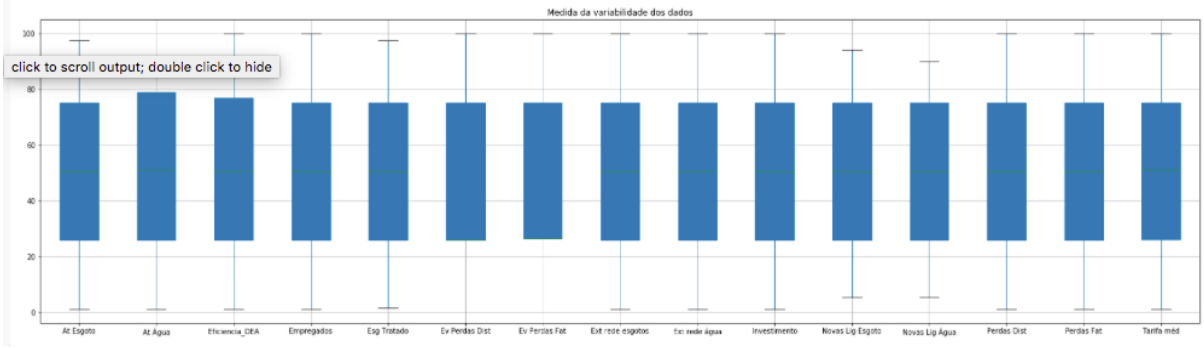
Graf7. *Gráfico da População pela Natureza.* As empresas SEMAPubl operam nas cidades mais populosas do Brasil.



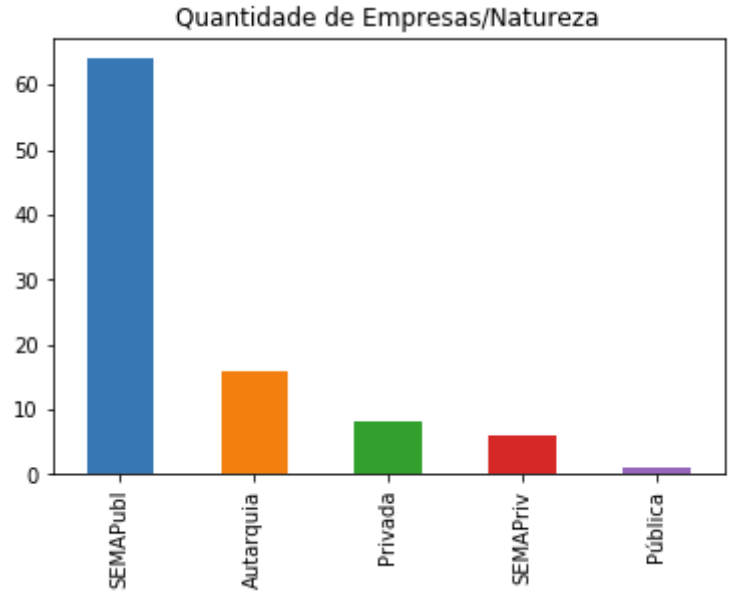
Graf8. *Gráfico da correlação entre os atributos.* Podemos identificar correlações entre: Novas Lig Esgoto/At Esgoto, Novas Lig água/At água, Ext rede água/esgoto e Perdas Dist/Fat



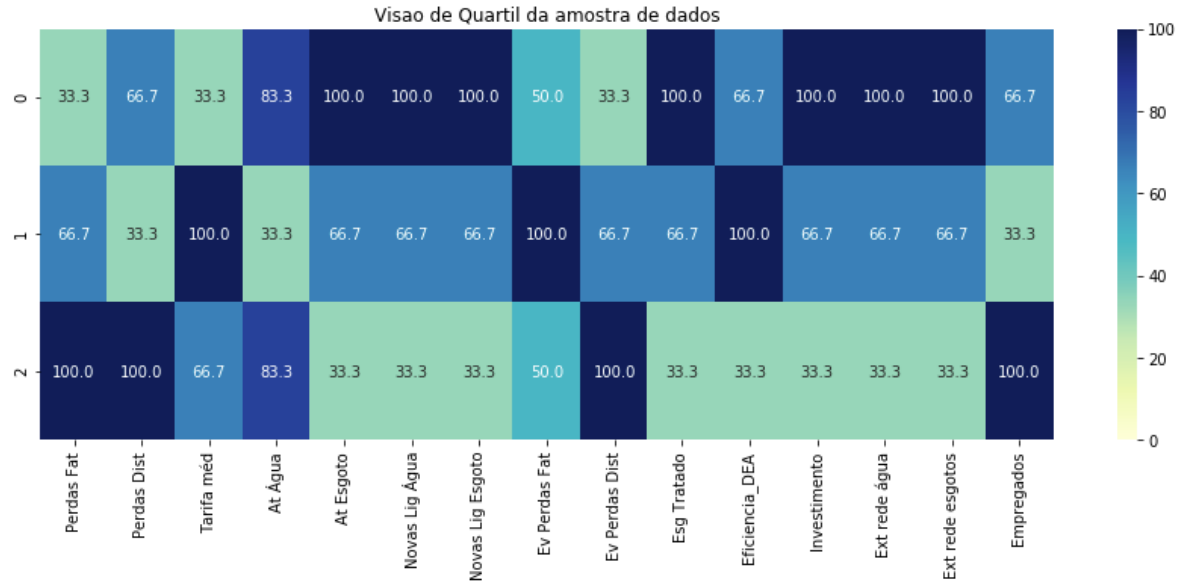
Graf9. *Gráfico da medida da variabilidade dos dados.* Em geral, todos os atributos possuem uma variabilidade equiparada.



Graf10. *Gráfico da quantidade de empresas por Natureza.* Mais da metade das empresas selecionadas são de Sociedade Econômica Mista de Administração Pública enquanto menos de 20 são privadas.



Graf11. *Visualização das amostras com relação à sua posição nos percentis*. A amostra(0) é líder, estando acima do quartil 3 nos índices de atendimento(esgoto, novas ligações de água/ligações faltantes, novas ligações de esgoto/ligações faltantes), esgoto tratado por água consumida. Também possui o maior investimento, bem como a maior extensão da rede de água e esgoto. Podemos considerar que seja a maior cidade do país. Essa amostra tem perdas faturamento, tarifa média e evolução nas perdas de distribuição abaixo do quartil 3, levando a considerar que tem um índice de perda baixo comparados com os índices de produção.
A amostra(1) tem uma tarifa média alta e uma evolução de perda de faturamento também é considerada alta, ambos estão acima do quartil 3. Tem a maior parte dos índices de atendimento no quartil 3. E seus investimentos e extensão de rede de água e esgoto são equivalentes aos índices de atendimento.
A amostra(2) tem alta perda nos índices de faturamento, distribuição, evolução nas perdas de distribuição. Possui muitos empregados. Mais da metade dos atributos estão abaixo do quartil 3.
Tendo em vista que todas as amostras receberam investimento possivelmente são empresas de economia mista ou privada.



Algoritmos e técnicas

Iremos explorar os dados por meio de visualizações e códigos para compreender a relação entre a natureza jurídica de cada DMU e determinar qual o tipo de empresa mais influencia o saneamento básico das cidades no Brasil. Também iremos observar estatística do conjunto de dados, levando em consideração a relevância de cada natureza jurídica, e selecionar alguns exemplos de pontos de dados do conjunto de dados para acompanhar durante o andamento do projeto.

De antemão, o conjunto de dados é compostos por sete natureza jurídica importantes de empresa: Administração pública direta, autarquia, empresa privada, empresa pública, organização social, sociedade de economia mista com administração privada, sociedade de economia mista com administração pública.

Iremos selecionar algumas amostras de dados de pontos e explorá-los com mais detalhes, isso vai facilitar a compreensão da análise das DMUs.

Para garantir significativamente a importância dos resultados obtidos iremos pré-processar os dados para representar as DMUs e realizar um escalonamento dos dados também como remover os dados aberrantes(outliers). Dados aberrantes podem enviesar resultados que levam em consideração os pontos de dados.

Podemos obter a rentabilidade dos acionistas pelo giro do ativo, lucro e endividamento de uma dada empresa. Essa rentabilidade serve para medir a eficiência das empresas. Porém, há áreas em que a rentabilidade aos acionistas(geração de valor aos acionistas) não é suficiente. Ou seja, há áreas em que o bom desempenho financeiro não é suficiente para saber se uma atividade está sendo eficiente diante de uma organização ou de uma sociedade. Os exemplos mais típicos disso são justamente as organizações públicas, pois não registram lucro. Essas organizações precisam mensurar a eficiência de qualquer forma. Charnes, Cooper e Rhodes em 1978 apresentaram um modelo matemático chamado DEA-CCR, é a junção de 'Data Envelopment Analysis', Análise Envoltória de Dados em português, com as iniciais de seus criadores. Presume-se que todas as unidades possuem um retorno constante de escala. O primeiro modelo DEA, avalia a eficiência entre unidades e identifica quais são as unidades(beachmarking) referência para as Unidades Tomadoras de Decisão(DMU - Decision Making Units) que não são eficientes. O DEA não precisa usar dados financeiros para medir a eficiência de um conjunto de empresas.

Então, os resultados serão decorrentes do cálculo da eficiência operacional de cada empresa a partir do DEA. O DEA-CCR vai dar como resultado um indicador de no máximo 1. A empresa que tiver DEA igual a 1 é porque foi 100% eficiente, já a empresa que tiver DEA igual a ou inferior a 0,99 quer dizer que foi dada proporção eficiente.

O modelo válida a proposta inicial de Charnes, Cooper e Rhodes (1978), onde as DMUs que, para demonstrar eficiência, devem usar o mínimo de insumos, ou inputs possíveis para constituir o máximo de produtos, ou outputs possíveis.

Para o setor de saneamento básico, há que empregar como variáveis de insumos e produtos itens semelhantes à proposta de Carmo (2003). Os insumos seriam a mão-de-obra empregada, a capacidade instalada, a extensão da rede de distribuição e a extensão da rede de coleta. Os produtos seriam o volume de água faturado, o volume de esgoto faturado, a economia ativa de água e a economia ativa de esgoto. A proposta é usar os dados do Sistema Nacional de Informações sobre Saneamento. Os outputs selecionados são: Indicador de atendimento urbano de água (%), Indicador de atendimento urbano de esgoto (%), Indicador de esgoto tratado por água consumida (%), Indicador novas ligações de água/ligações faltantes (%), Indicador novas ligações de esgoto/ligações faltantes (%), Indicador evolução nas perdas de faturamento (%), Indicador evolução nas perdas de distribuição (%); e como inputs: Indicador perdas no faturamento 2016 (%), Indicador perdas na distribuição 2016 (%),

Tarifa média (R$/m³), LN Investimento, LN extensão rede de água, LN extensão rede de esgotos, LN do nº de funcionários em 2016.

Há um problema chamado "problema do tamanho" em usar dados absolutos como por exemplo a extensão da rede de água ou de esgoto. Imagina compara Franca, que é uma cidade pequena e verticalizada, uns 500 metros de canos já basta para ter saneamento eficiente, com o Rio de Janeiro, que é uma cidade muito populosa e horizontalizada, deve precisar de uns 12 mil km de encanamento e ainda não basta. Por isso não é bom usar dados absolutos nas análises quantitativas. A média e a mediana variam significativamente (indicando um grande desvio), portanto vamos aplicar um escalonamento não linear. Vamos reduzir o desvio aplicando o algoritmo natural. Os dados que tratamos com logaritmo natural estão indicados com as iniciais "LN". Após aplicar o algoritmo natural para o escalonamento dos dados, a distribuição para cada atributo deve parecer mais normalizado.

Vamos usar a análise de componentes principais (PCA) para elaborar conclusões sobre a estrutura subjacente de dados de DMUs. A PCA calcula as dimensões que melhor maximizam a variância das natureza jurídica envolvidas, encontrando as combinações de componentes de natureza jurídica descrevem as DMUs.

Criar um método de pontuação simples de um dado clustering para identificar o número ótimo de cluster melhor segmenta os dados.

Por fim, utilizando o modelo de clusterização escolhido, iremos calibrar para obter os melhores hiperparâmetros e documentá-lo.

Benchmark

O DEA permite, entre outros aspectos relevantes, avaliar a eficiência das empresas que sejam referência(benchamarking) para as demais empresas analisadas que não possuem um desempenho 100%. A empresa com desempenho mais alto vai encapsular todas as empresas. Reparando que uma regressão não permite fazer isso. Não tem como avaliar a eficiência por meio de uma análise de regressão. Então, os resultados do índice DEA comporão a variável dependente para a análise de estatística inferencial que permitirá analisar o relacionamento entre a estrutura de propriedade e a eficiência operacional das empresas de saneamento básico das principais cidades brasileiras.

III. Metodologia

Pré-processamento de dados

Nesta seção, iremos pré-processar os dados para criar uma melhor representação das empresa ao executar um escalonamento dos dados. Detectando os discrepantes para saber que decisão tomar a respeito deles, assegurando que os resultados obtidos na análise são importantes e significativos.

Vamos reduzir o desvio aplicando o algoritmo natural nos atributos: Investimento, Ext rede água, Ext rede esgotos e Empregados. Após aplicar o algoritmo natural a distribuição para cada atributo deve parecer mais normalizado.

Implementação

Implementação: Escalonando Atributos

A padronização de um conjunto de dados para uma unidade comum entre os atributos é muito importante para estimar o aprendizado de máquina: eles podem arbitrar mal se o atributo individual não for da mesma unidade de medida dos dados normalizados distribuídos (por exemplo, Gaussiano com média 0 e variação unitária).

Na implementação do escalonamento usamos a biblioteca StandardScaler[10] do sklearn para padronizar os atributos removendo a média e o dimensionamento para a variação da unidade.

A centralização e o dimensionamento são independentes para cada atributo, calculando as estatísticas relevantes das amostras no conjunto de treinamento. Então, obtemos a média e o desvio padrão para serem usados no método de transformação.

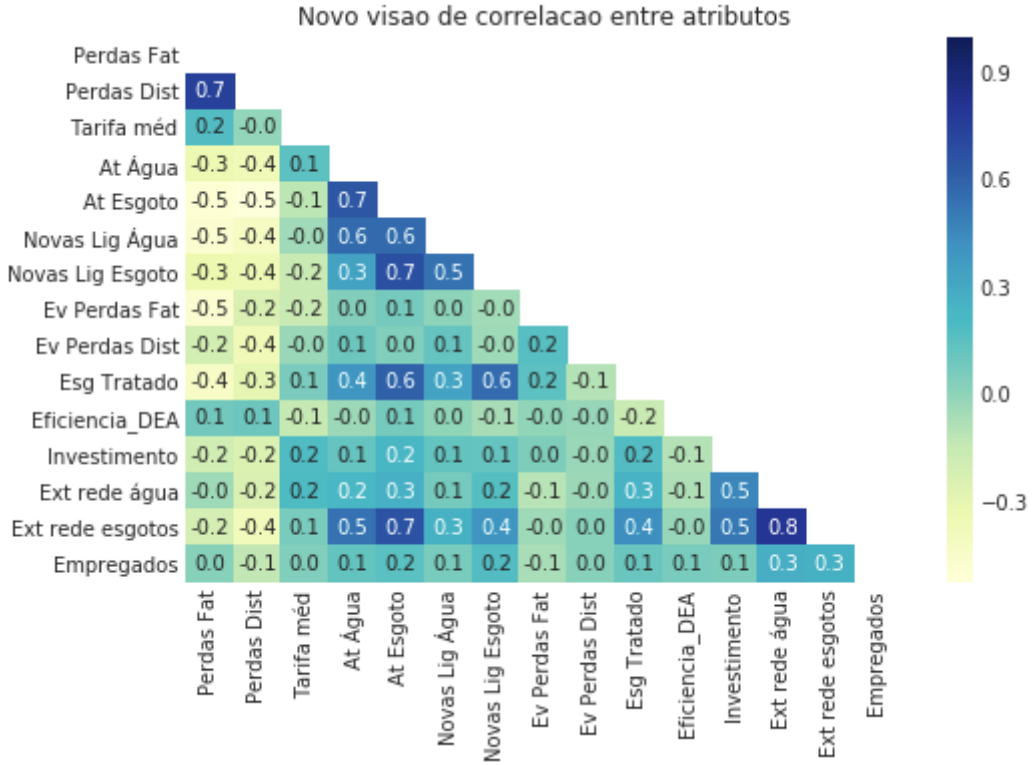
O escalonamento transforma os dados, fazendo com que a variância de cada atributo seja unitária para o modelo performar melhor.

[10] "sklearn.preprocessing.StandardScaler — scikit-learn 0.19.1 ...." http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Acessado em 23 jun. 2018.

Tab5. *Tabela da amostra de dados escalonada.*

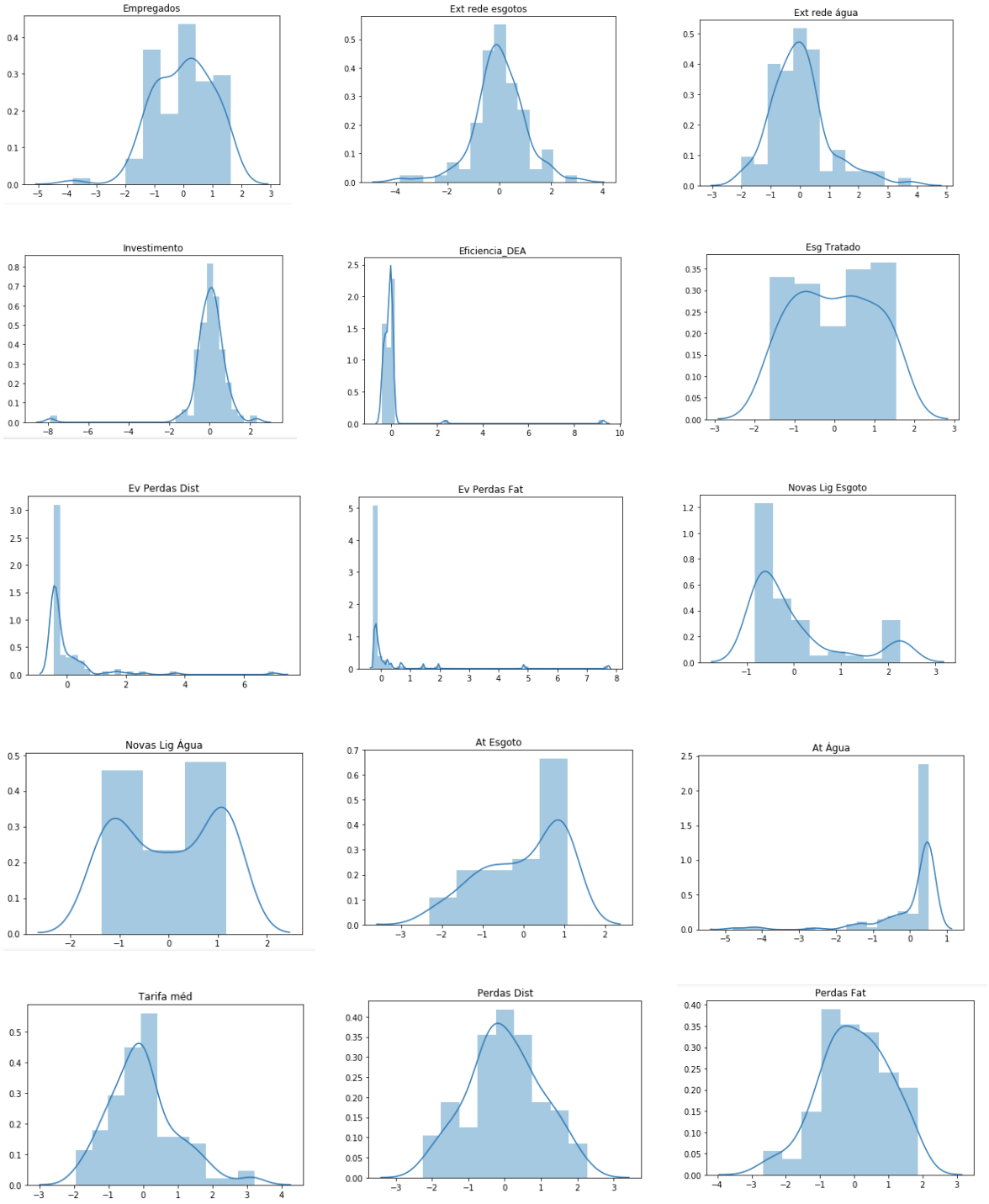
	Perdas Fat	Perdas Dist	Tarifa méd	At Água	At Esgoto	Novas Lig Água	Novas Lig Esgoto	Ev Perdas Fat	Ev Perdas Dist	Esg Tratado	Eficiencia_DEA	Investimento	Ext rede água	Ext rede esgotos
0	-0.600607	-0.403459	0.091762	0.506916	0.986121	0.604162	0.940444	-0.262217	-0.437521	0.344721	-0.319420	2.325160	3.789932	3.003657
1	0.915040	-1.290252	0.607582	0.431364	0.580661	0.129712	-0.294058	-0.246659	-0.018069	-0.209133	-0.309639	1.458984	2.838261	1.814404
2	1.170928	1.359953	0.225875	0.506916	-0.545998	-0.040263	-0.810791	-0.262217	0.317493	-0.641108	-0.338228	0.794442	0.757702	0.223522

Graf11. *Visualização da correlação dos atributos escalonados.* As mesmas correlações permanecem e aparecem novas boas correlações: At esgoto/At água e Ext rede esgoto/At esgoto.



Após aplicar o algoritmo natural para o escalonamento dos dados, a distribuição para cada atributo deve parecer mais normalizado. Para muitos pares de atributos, observe nos próximos gráficos essa correlação.

Graf12. *Visualização centralizada da escala média/atributo para a variância unitária.*

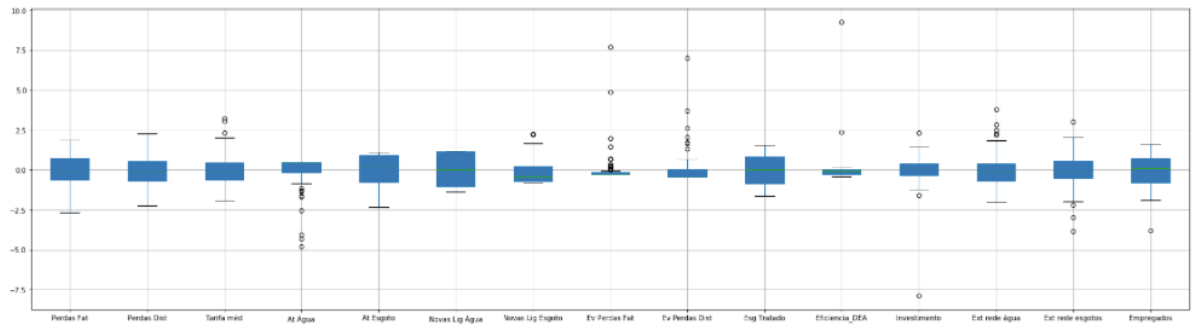


Implementação: Detecção de Discrepantes

Utilizamos a biblioteca numpy.percentile[11] para implementar o código e obter o cálculo percentil dos quartis dos dados.

Atribuímos o valor do 25º percentil do atributo dado para o Q1. Atribuímos o valor do 75º percentil do atributo dado para o Q3. Atribuímos o cálculo de um passo do discrepante do atributo dado para o step. Como é opcional, não removemos os pontos de dados do conjunto de dados considerados discrepantes. Os outliers viesam o resultado da análise e o seu comportamento é parte dessa análise. Portanto não serão excluídos. Portanto alguns indicadores terão uma variação muito grande nos dados das empresas, fazendo com que haja scores muito altas ou muito baixas para esses indicadores.

Graf13. *Visualização dos dados Discrepantes.* No conjunto de dados não temos discrepantes duplicados. A correlação entre as categorias com os outliers e sem os outliers têm resultados semelhantes, não há diferença significativa em removê-los. Depois que os atributos foram escalonados, a amplitude entre a média e a mediana diminuiu se comparada com os originais. O desvio padrão também diminuiu como podemos ver no distplot acima. Os quartis também ficaram muito próximos. Optamos por não remover os discrepantes, que são quase a metade dos registros podendo ocasionar uma perda grande de informação. Acredito que o viés que eles causam é parte da influência dessa análise.



[11] "numpy.percentile - Numpy and Scipy Documentation - SciPy.org." https://docs.scipy.org/doc/numpy/reference/generated/numpy.percentile.html. Acessado em 23 jun. 2018.

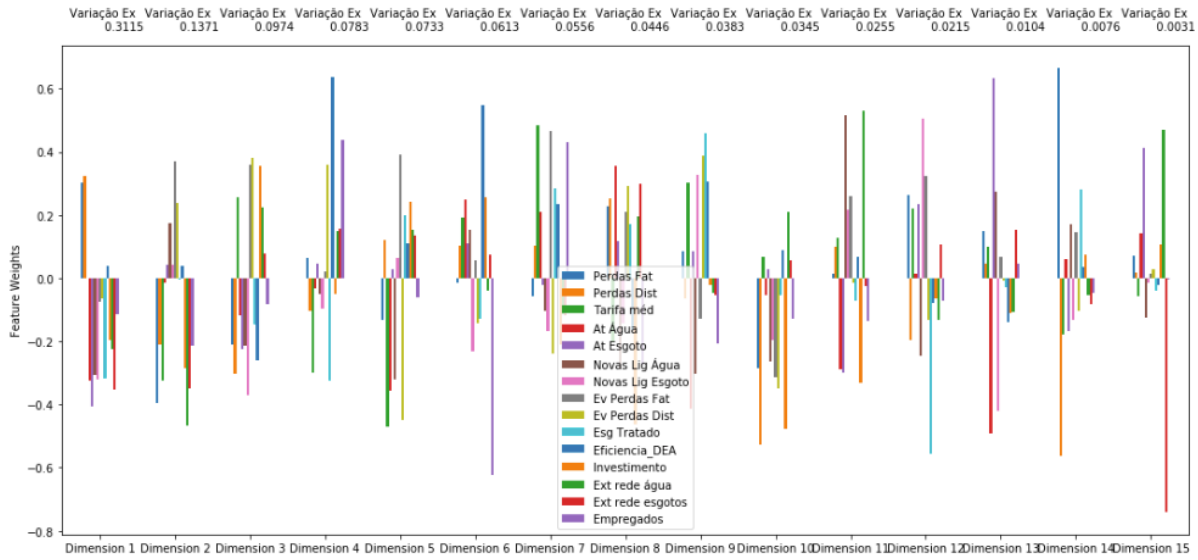
Refinamento

Transformação de Atributo

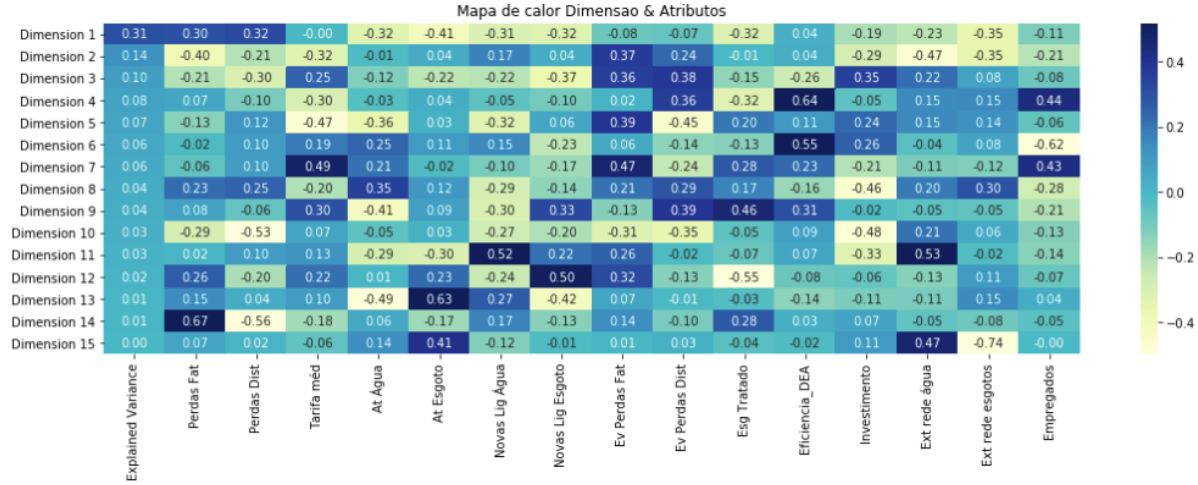
Nesta seção, vamos utilizar a análise de componentes principais (PCA) para elaborar conclusões sobre a estrutura subjacente de dados das empresas. Dado que ao utilizar a PCA em conjunto de dados calcula as dimensões que melhor maximizam a variância, nós iremos encontrar quais combinações de componentes de atributos melhor descrevem as empresas.

Agora iremos aplicar a PCA no conjunto de dados para descobrir qual dimensão e dos dados melhor maximizam a variância dos atributos envolvidos e a razão da variância explicada de cada dimensão.

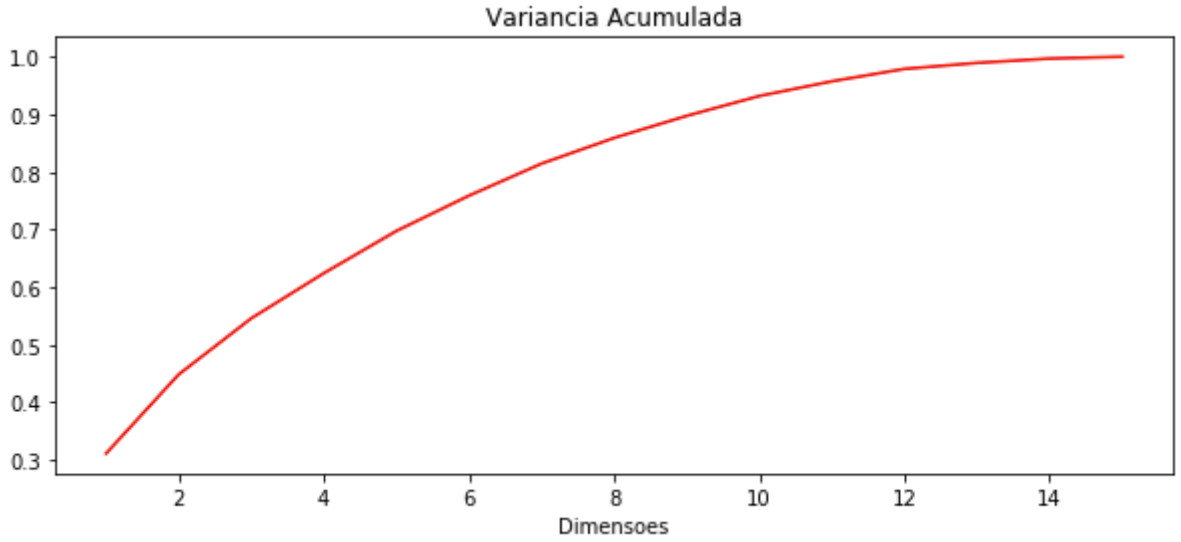
Graf14. *Gráfico de variância explicada para 15 dimensões.* Aplica a PCA ao ajustar os dados com o número de dimensões igual ao número de atributos. Transforma a amostra de df_scaled_data_samples utilizando o ajuste da PCA acima. Gera o plot dos resultados da PCA. Para verificar a variância explicada pelas 15 componentes. A dimensão 2 é fortemente correlacionada com Novas Lig Água e Esgoto, Ev Perdas Fat e Dist, Esg Tratado e Eficiência atingindo 45% da variância junto com a dimensão 1. A dimensão 8 a variância acumulada é de 86%. Sendo bem correlacionada com Perdas Fat, Perda Dist, At Água, At Esg, Ev Perdas Fat, Ev Perdas Dist, Esg Tratado, Ext rede água e esgoto.



Graf15. Visualização do *mapa de calor entre 15 Dimensões e os Atributos dos dados.* A dimensão 1 aumenta com 3 atributos. Se correlaciona com Eficiência, mas principalmente com Perdas Fat, Perdas Dist. Explicando 31% da variância no conjunto de dados.



Graf16. *Gráfico de variância acumulada.* Analisando as 8 primeiras dimensões, a Ext Rede Água e Ext Rede Esgoto mostram uma leve tendência linear e é o par mais correlacionado do conjunto de dados com 0.96 de coeficiente e tem um viés pequeno.

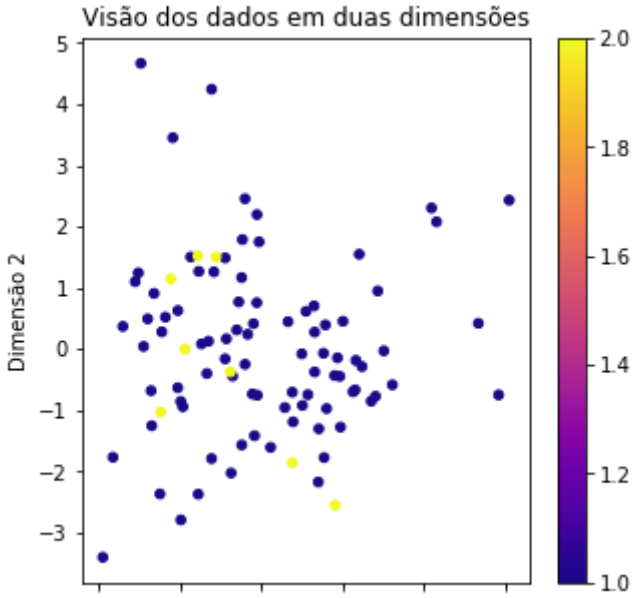


Tab6. *Tabela das amostra depois de aplicado a transformação da PCA*

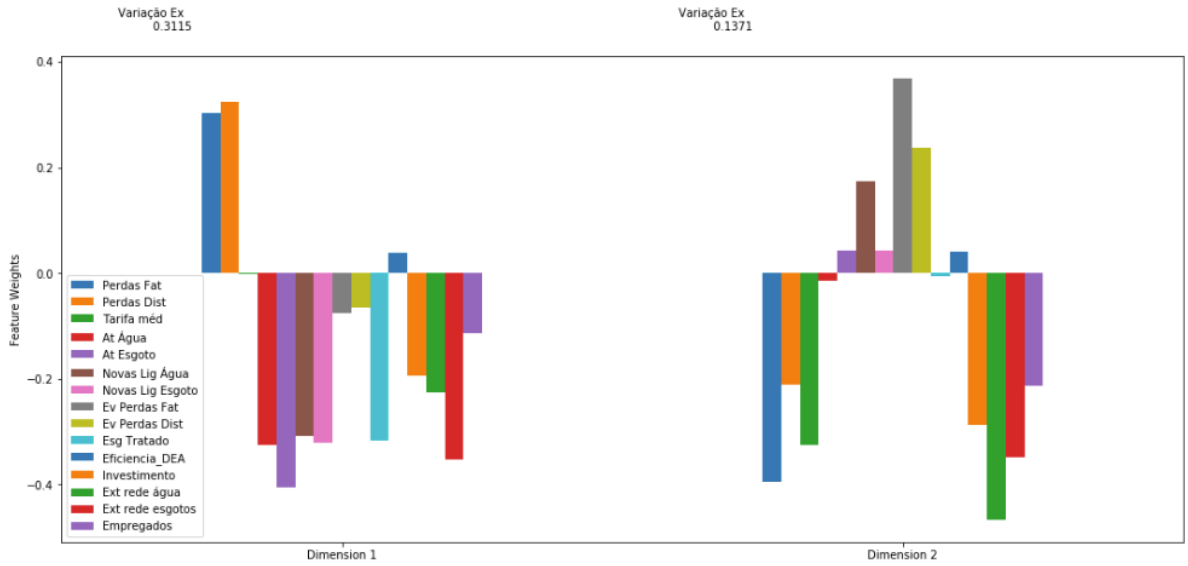
	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6	Dimension 7	Dimension 8	Dimension 9	Dimension 10	Dimension 11	Dimension 12	Dimension 13	Dimension 14	Dimension 15
0	-3.8976	-3.4016	1.1224	0.6826	1.3266	0.0800	-1.0009	-0.0060	-0.6423	-0.0086	0.9808	-0.0768	-0.1092	-0.5749	0.1116
1	-1.9783	-2.7909	1.5421	0.6437	0.0923	0.3217	-0.7488	0.2981	-0.4056	0.5207	0.5448	0.3613	0.3035	0.9062	0.4528
2	0.7632	-1.8570	0.3488	0.6791	-0.4828	-0.3532	0.2524	0.2911	-0.9980	-1.3022	-0.0844	-0.3914	-0.0405	-0.1245	0.2639

Implementação: Redução da Dimensionalidade

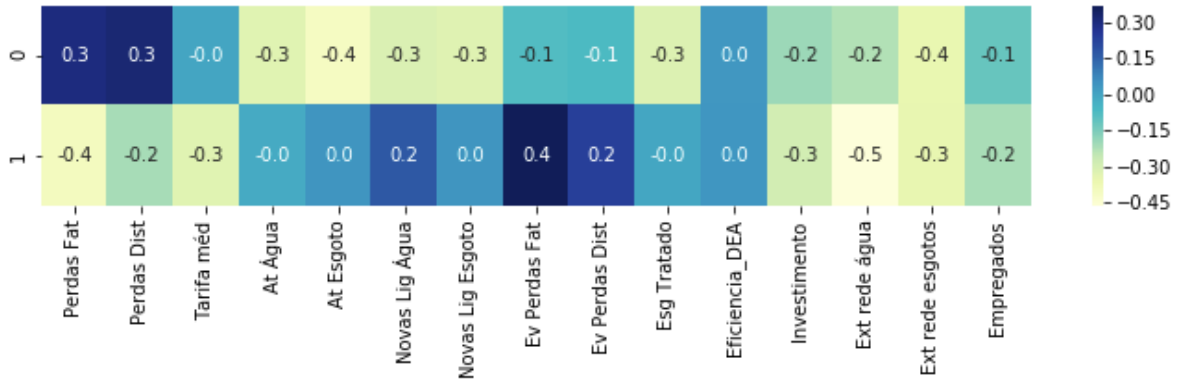
Graf17. *Visão dos dados Dimensão 1/Dimensão 2*. Reduzir a dimensionalidade dos dados e reduzindo a complexidade do problema. O espaço dimensional de 15 dimensões para 2 dimensões não separar muito bem os dados visualmente.



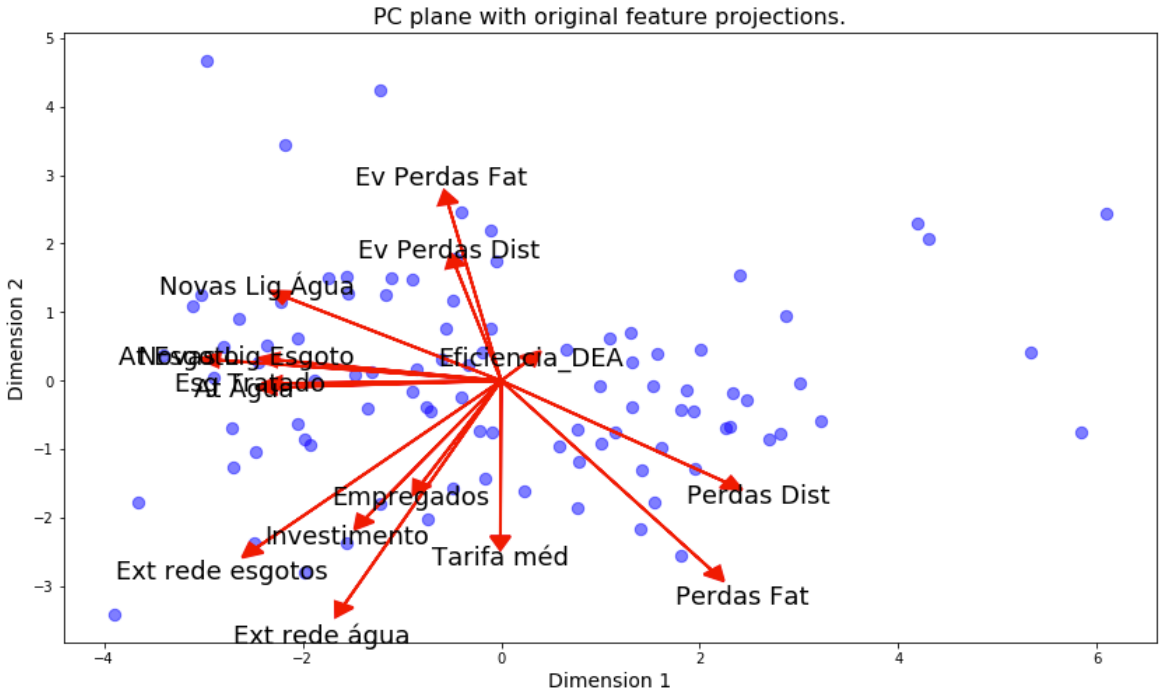
Graf18. *Gráfico de variância explicada para duas dimensões.*



Graf19. Visualização do *mapa de calor entre duas Dimensões e os Atributos dos dados.* Os valores das duas primeiras dimensões permanecem iguais quando comparados com a transformação do PCA em quinze dimensões.



Graf20. *Visualização de dispersão.* Cada ponto é representado por sua pontuação junto dos componentes principais. Os eixos Dimension 1 e Dimension 2 são os componentes principais. Podemos ver a projeção dos atributos originais junto dos componentes. Assim podemos interpretar a redução da dimensionalidade dos dados e descobrir relacionamentos entre as componentes principais e os atributos originais. Os atributos correlacionados entre si são: Ev Perdas Fat/Ev Perdas Dist, At Água/At Esgoto/Esg, Tratado/Novas Lig Esgoto e Empregados/Investimento/Est rede Água e Esgoto.

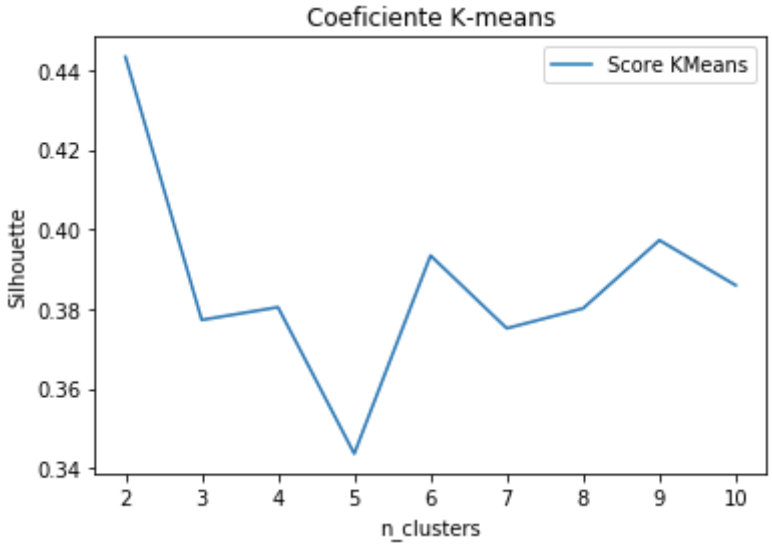


Aplicamos o algoritmo de cluster K-means por ter custo computacional mais baixo, é mais rápido e mais eficiente para esse conjunto de dados. O K-means distingue rigidamente os limites de um cluster do outro. O K-means não diverge os locais dos pontos, dimensionando bem o conjunto de dados. O algoritmo tem o custo de execução mais baixo que o GMM, para extração dos clusters. Também calcula somente a média dos pontos atribuídos a um cluster.

Tab7. *Tabela de coeficientes clusters*

	Score KMeans
2	0.443407
3	0.377277
4	0.380476
5	0.343657
6	0.393456
7	0.375162
8	0.380187
9	0.397311
10	0.385993

Graf21. *Visualização dos coeficientes K-means.* O melhor coeficiente foi para 2 clusters ficando: 2 clusters = 0.44.



IV. Resultados

As consideráveis deficiências existentes na provisão de serviços, a baixa cobertura, os elevados índices de ineficiência e os grandes montantes de investimento aplicados botaram em questão o desempenho das empresas públicas e privadas. As comparações referem-se a vários aspectos: cobertura da rede de saneamento, atendimento, novas ligações, tratamento de esgoto e outros.

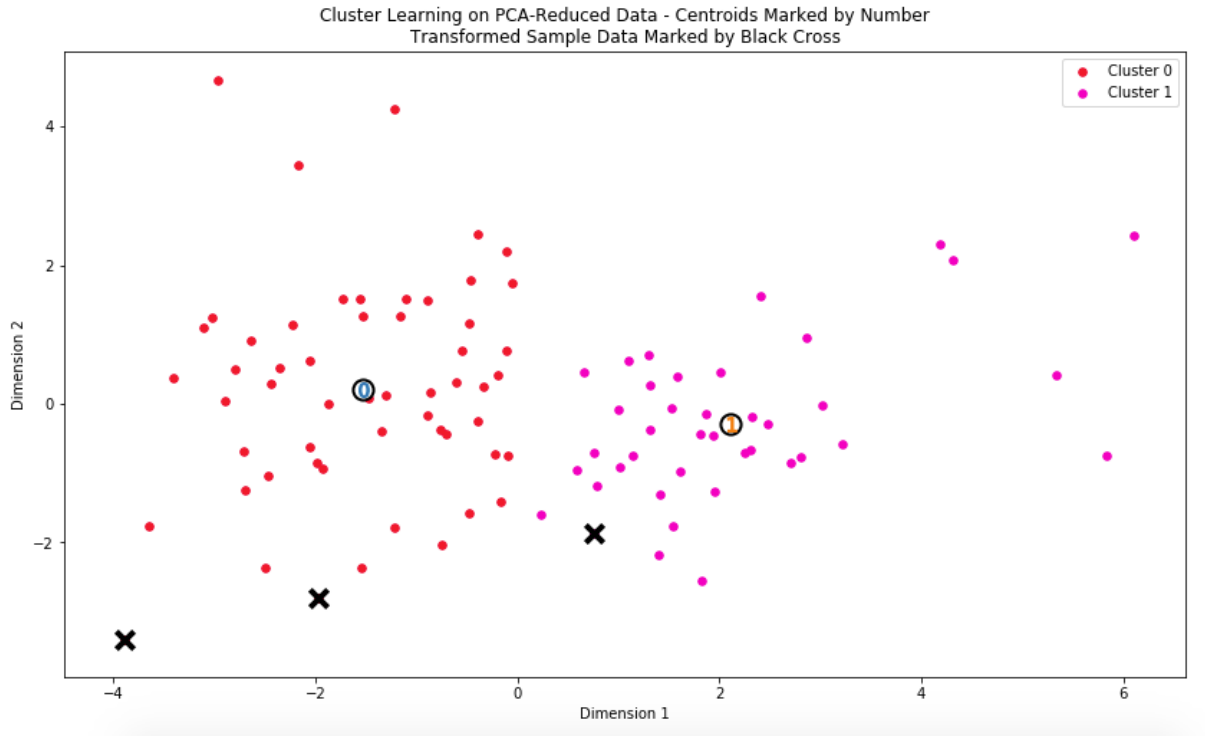
As empresas privadas evidenciam menores índices de perdas de distribuição e faturamento que as empresas públicas, sendo um bom sinal de eficiência. Eficiência que pode ser apontada, por exemplo, pela utilização de uma tecnologia mais avançada. Por outro lado as empresas públicas tiveram desempenho favorável em vários indicadores: atendimento de água e esgoto, perdas na distribuição, despesa com os serviços e tarifa média praticada.

As principais diferenças das empresas públicas e privadas se pode compreender que não seria tanto teórico, mas sim embutida na finalidade da adoção das boas práticas inerentes ao conceito. O objetivo no setor privado ficaria na busca pelo resultado, e no setor público, a busca pela conformidade.

Os resultados parciais indicam que a maior parte é de empresas públicas, sendo em torno de 70% das empresas. Na amostra de dados as amostras (0) e (1) são empresas públicas.

Modelo de avaliação e validação

Graf22. *Visualização dos clusters.* Os dados de amostra estão marcados em **x**, e estão segregados em seus respectivos clusters que foram aprendizados no modelo de clusters K-means aplicados no PCA dos dados.



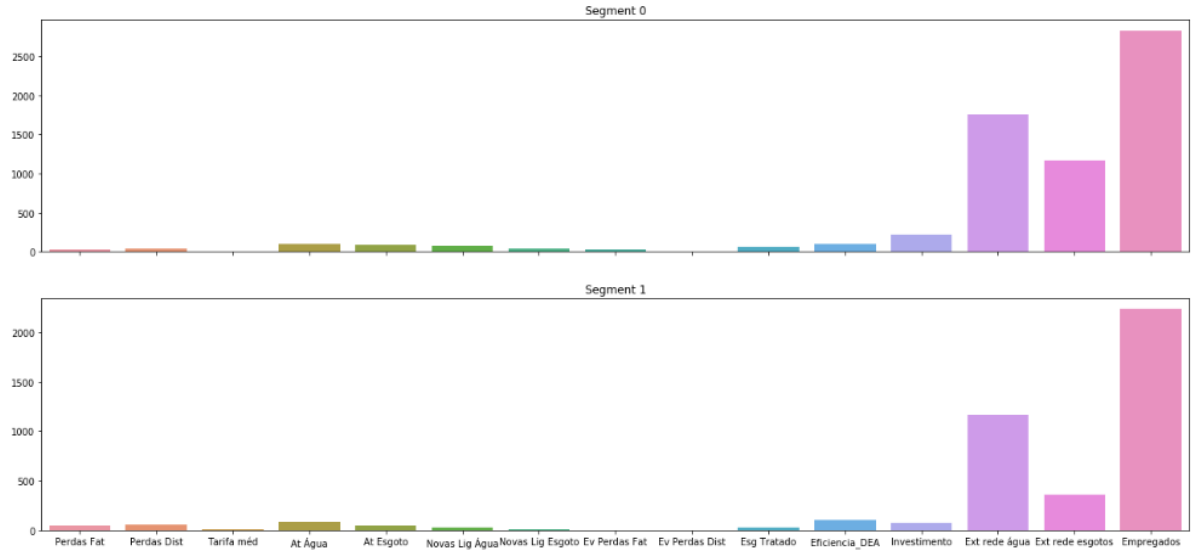
Implementação: Recuperação de Dados

Já que os dados foram atualmente reduzidos em dimensões e escalas por um algoritmo, nós podemos recuperar os atributos representativo das empresas desses pontos de dados ao aplicar transformações inversas para observar a eficiência do modelo.

Tab8. *Tabela com os centros do conjunto de dados original.*

	Perdas Fat	Perdas Dist	Tarifa méd	At Água	At Esgoto	Novas Lig Água	Novas Lig Esgoto	Ev Perdas Fat	Ev Perdas Dist	Esg Tratado	Eficiencia_DEA	Investimento	Ext rede água	Ext rede esgotos	Empregados
Segment 0	26.0	35.0	4.0	100.0	87.0	74.0	43.0	25.0	5.0	66.0	94.0	214.0	1762.0	1170.0	2826.0
Segment 1	50.0	51.0	4.0	84.0	43.0	26.0	4.0	-0.0	2.0	30.0	107.0	78.0	1167.0	363.0	2231.0

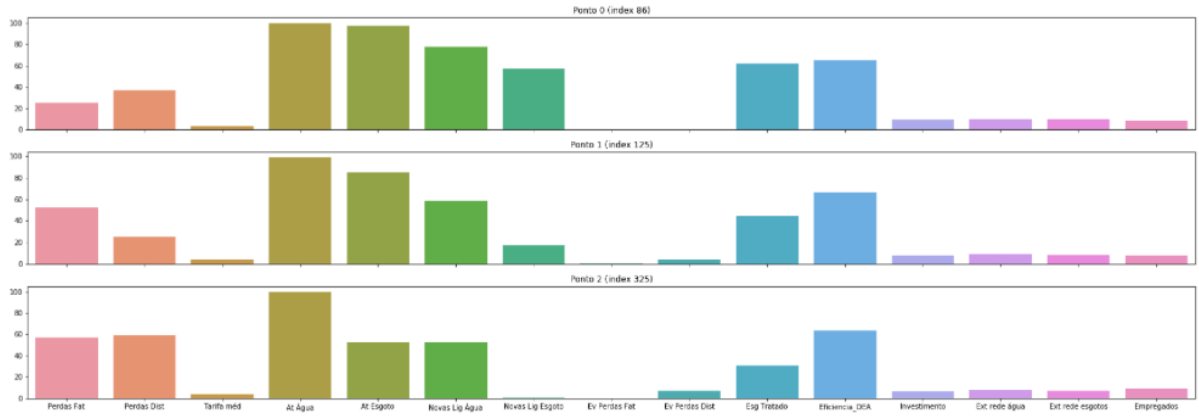
Graf23. *Visualização Representação de segmento das empresas.* Considerando os totais de cada atributo, os centros de clusters representam dois tipos de empresas características muito semelhantes, não sendo viável a identificação dos mais eficiênte em seus respectivos segmentos, nem por suas características. Cada atributo é ligeiramente igual nos dois segmentos. Essas características podem ter sido herdadas pela estrutura da Natureza Jurídica que acabam misturando empresas públicas com administração privada, o que confunde a identificação de quais tipos de empresas seriam mais eficientes.



Representação de segmento da amostra de empresas

A amostra (0) foi prevista para o cluster (0), amostra (1) prevista para o cluster (0) e a amostra (2) prevista para o cluster (1).

Graf24. *Visualização Representação de segmento das empresas.* Com observações semelhantes ao gráfico anterior com diferenças sutis. O algoritmo conseguiu separar os segmentos com bom desempenho, como iremos constatar na conclusão do projeto, empresas do segmento 0 são públicas e do segmento 1 são privadas.



Justificativa

Consideramos dois segmentos de empresa(pública e privada). Por fim, comparamos os segmentos de empresas com uma variável o atributo 'Tipo' no conjunto de dados, para ver se o cluster identificou corretamente a relação.

Temos dois grandes segmentos previstos pelo modelo que são referentes às tipos de empresas de saneamento no Brasil.

Um fator curioso para definir as empresas é que pelas características que atribuímos, os dois segmentos são semelhantes em desempenho.

Uma aprendizagem não supervisionada pode ser utilizada para treinar as empresas dos dados originais. Suas predições podem ser usado para prever o segmento de clientes para novas empresas, possibilitando a escolha do tipo de empresa mais apropriado para saneamento.

V. Conclusão

Visualização de forma livre

Uma variável oculta é a saúde da população. De maneira que os serviços estejam sendo realizados adequadamente a população tende a ser mais saudável. Não faz sentido ter uma população com muitas doenças se tem um serviço de saneamento adequado, gerando prejuízo ao população.

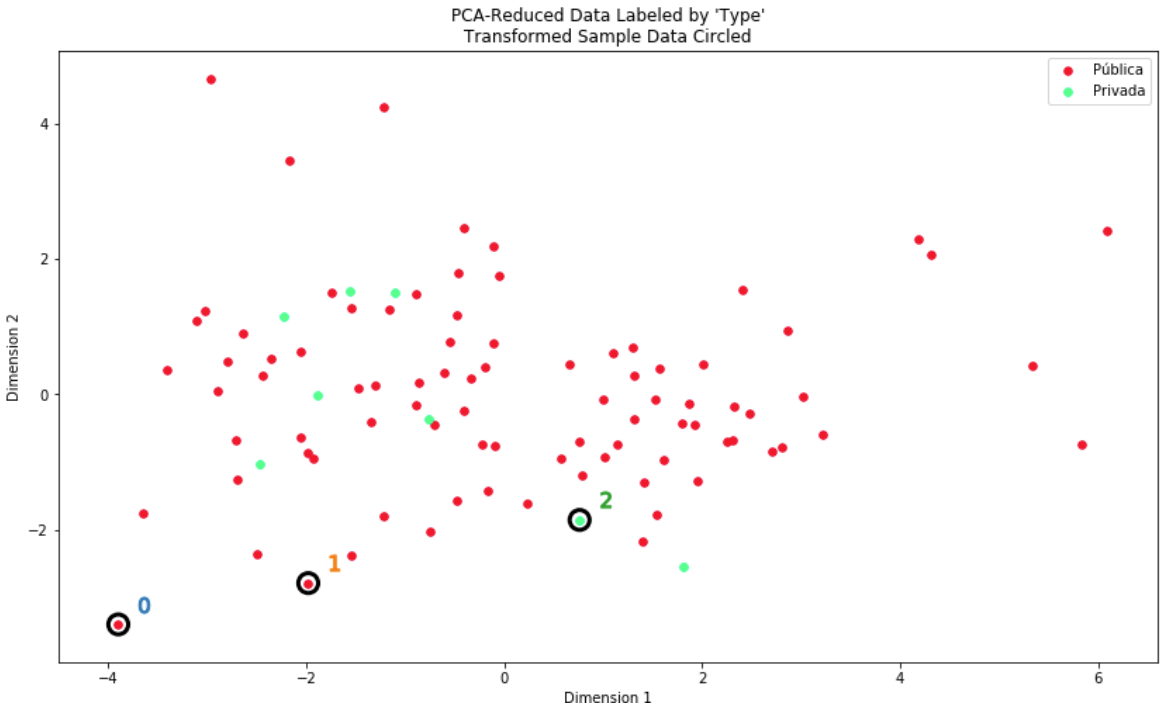
Proponho a realização de testes A/B nos serviços das empresas e medir a satisfação pelo serviços para evitar qualquer tipo de coincidência que possam ocorrer na análise. Isso nos dará informações mais confiáveis e ampliará nosso conjunto de dados.

Ao agrupar as empresas, temos uma melhor ideia de qual segmento de empresa pode ser mais eficiente e aplicar mudanças nas prefeituras para que os resultados sejam mais precisos. Mas não podemos considerar um grupo mais eficiente que o outro nesse conjunto de dados.

Visualizando Distribuições Subjacentes

No começo deste projeto, não consideramos o atributo 'Tipo de empresa', então os atributos das empresas seriam enfatizadas na análise. Ao reintroduzir o atributo 'Tipo de empresa' ao conjunto de dados, uma estrutura interessante surge quando consideramos a mesma redução de dimensionalidade da PCA aplicada anteriormente no conjunto de dados original.

Graf25. *Visualização dos clusters.* Rotulados o PCA por Tipo de empresa das amostras transformadas estão circuladas e conferem com a predição do modelo proposto.



O modelo K-means separa os três círculos indicando os pontos da amostra escolhida. A distribuição se relaciona bem com o clusters previstos no modelo.

Os pontos (0) e (1) confirmam ser empresas públicas anteriormente como cluster (0) e o ponto (2) uma empresa privada, portanto confirma a previsão do modelo proposto, mostrando que o algoritmo K-means foi bom na distribuição das empresas.

Considero essa classificação sólida comparada com a definição de empresas obtive anteriormente.

Reflexão

A princípio exploramos os dados por meio de visualizações e códigos para compreender e determinar qual o tipo de empresa mais influencia o saneamento básico das cidades brasileiras. Vimos também estatística do conjunto de dados, levando em consideração a relevância dos atributos, e selecionamos 3 amostras do conjunto de dados para acompanhar durante o andamento do projeto e explorá-las com mais detalhes.

Em seguida, realizamos o pré-processamento dos dados para para obter uma representação das empresas e realizar um escalonamento dos dados e observamos a necessidade de remover os outliers. Consideremos permanecer os outliers para não perder informações que podem ser importantes.

Resolvemos o "problema do tamanho" de dados absolutos como por exemplo: Investimento, empregados e extensão da rede de água e esgoto aplicando logaritmo natural. Após aplicar o algoritmo natural aplicamos o escalonamento dos dados, a distribuição para cada atributo deve ficou mais normalizado.

implementamos a Análise de Componentes Principais (PCA) para elaborar conclusões sobre a estrutura subjacente de dados de empresas. A PCA calculou as dimensões que melhor maximizam a variância da empresas envolvidas, encontrando as combinações de componentes melhor descrevem as empresas.

Para identificar o número ótimo de cluster melhor segmenta os dados criamos um modelo de clusterização baseado em K-means para pontuar um dado clustering.

Por fim, utilizando o modelo de clusterização, calibramos a transformação do conjunto de dados para obter os melhores hiperparâmetros e documentá-lo.

Melhorias

Para ampliar o projeto, poderíamos adicionar mais atributos do conjunto de dados da NSIS, com o objetivo de obter novas perspectivas do novo conjunto de dados.

Também aplicaria um filtro mais refinado nos outliers de forma a detectar inconsistências relacionadas a erros de processo da empresa, como: se a empresa pública, porque está recebendo investimento? Por que as empresas privadas possuem muitos funcionários se temos um grande investimento? e outros.

Mas a principal mudança a ser aplicada é tratar o projeto todo pela natureza de jurídica. Isso demandaria muito tempo para eu fazer o que ultrapassa a entrega do projeto final.