RELATÓRIO DE CASE PARA RECURSOS HUMANOS

Guilherme Giuliano Nicolau
Cientista de Dados
PhD em Ciência Política (USP)

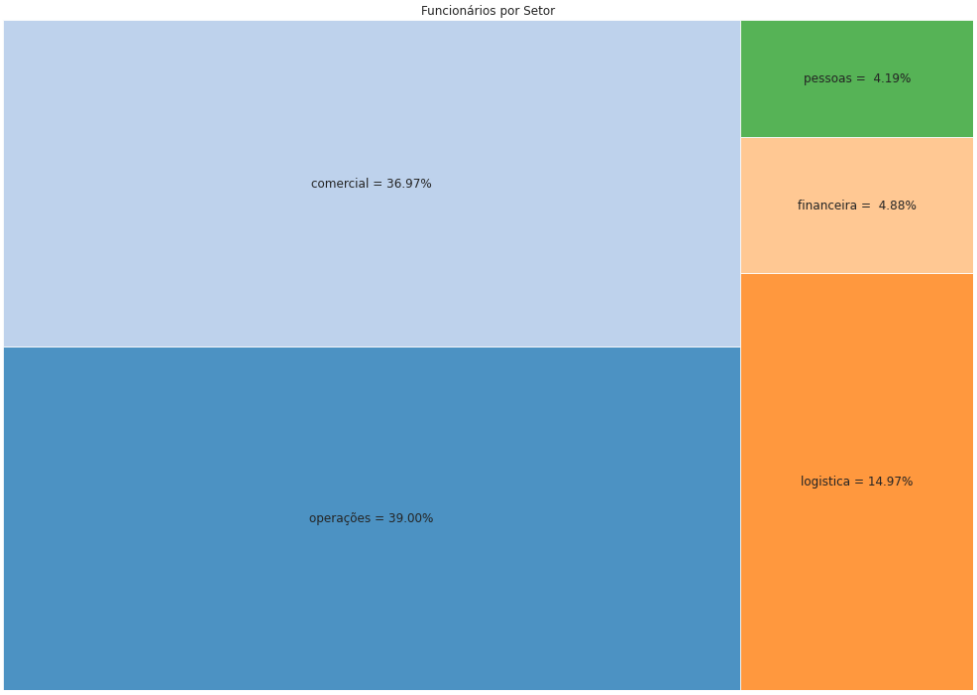
INTRODUÇÃO

Nosso **problema de negócio** é sobre a necessidade de melhorar a eficiência dos recursos humanos de uma empresa específica usando abordagens quantitativas e computacionais. Para isso, foram coletados dados psicométricos dos funcionários de diferentes áreas da empresa, além da avaliação de suas performances individuais a cada semestre.

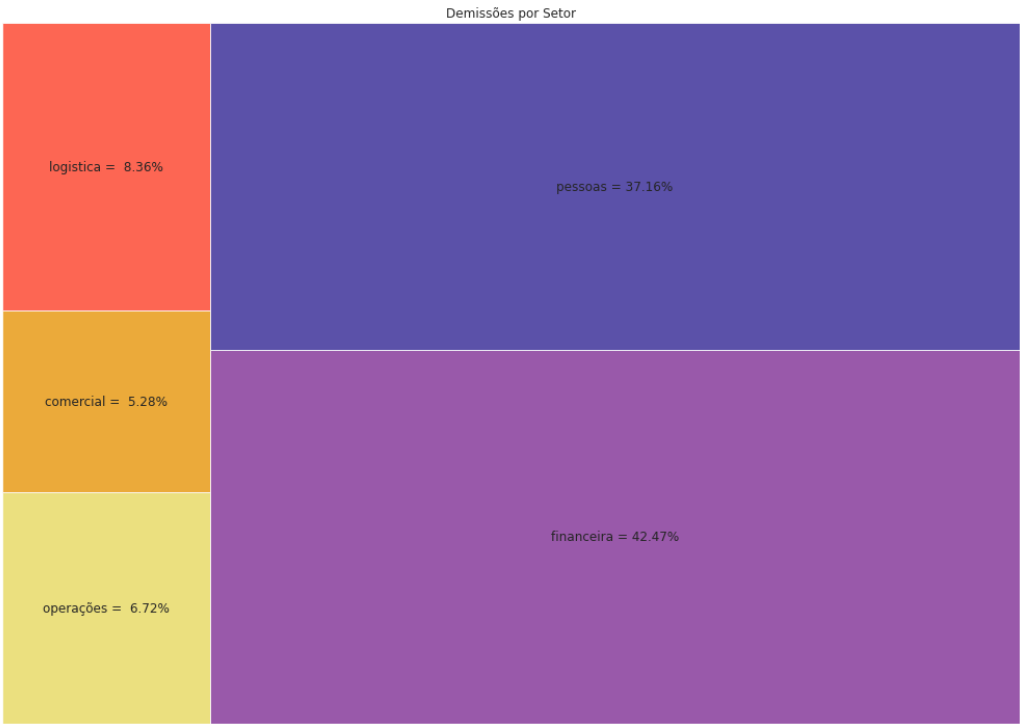
O **objetivo** é encontrar como podemos suprir as deficiências em recursos humanos para cada setor. Para isso, nossa **hipótese** é que podemos encontrar funcionários com habilidades subutilizadas em algum setor e realocá-lo onde possa ser reaproveitado, principalmente nos setores mais urgentes identificados. Para isso, optamos por **métodos** computacionais em ciência de dados, principalmente através de ferramentas como a linguagem de programação *Python*.

PARTE 1: Funcionários por setor e contratações

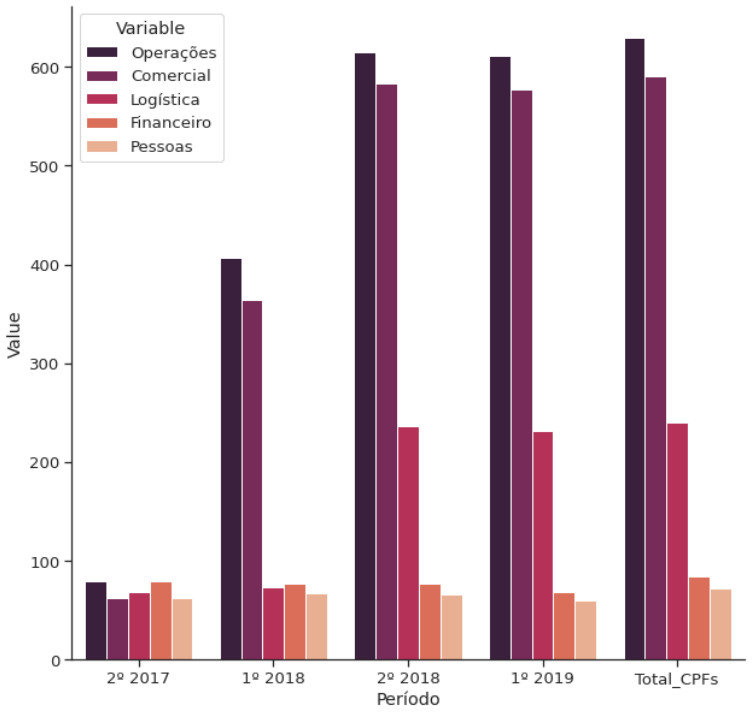
1) Até hoje passaram *1617 funcionários* (CPFs) na empresa, dividido pelas seguintes áreas:



2) Os setores que menos conseguem reter funcionários são *Pessoas* e *Financeiro* (quase metade das pessoas pessoas saíram do setor ou da empresa):

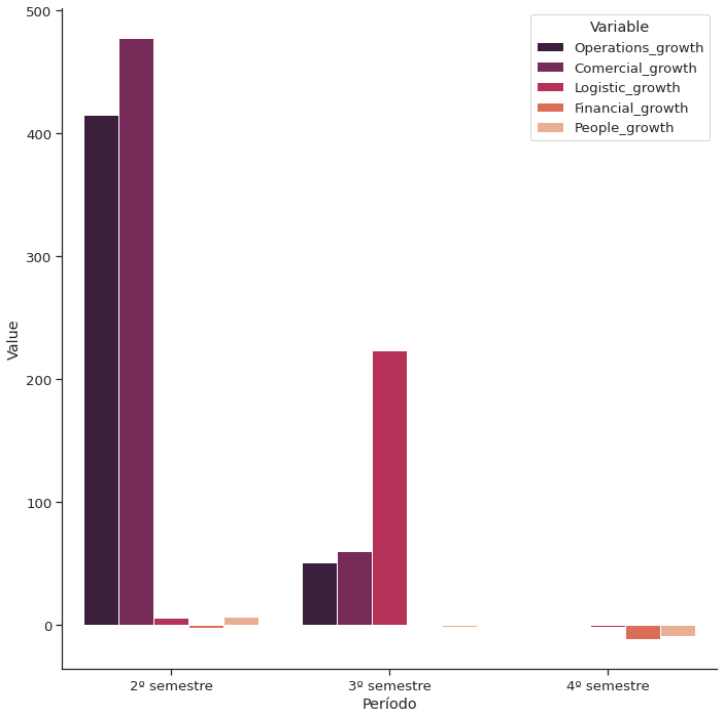


3) Ao longo do tempo, foram contratados muito mais funcionários das áreas de operações, comercial e também, um pouco menos, no setor de logística; o restante permaneceu mais ou menos no mesmo patamar.

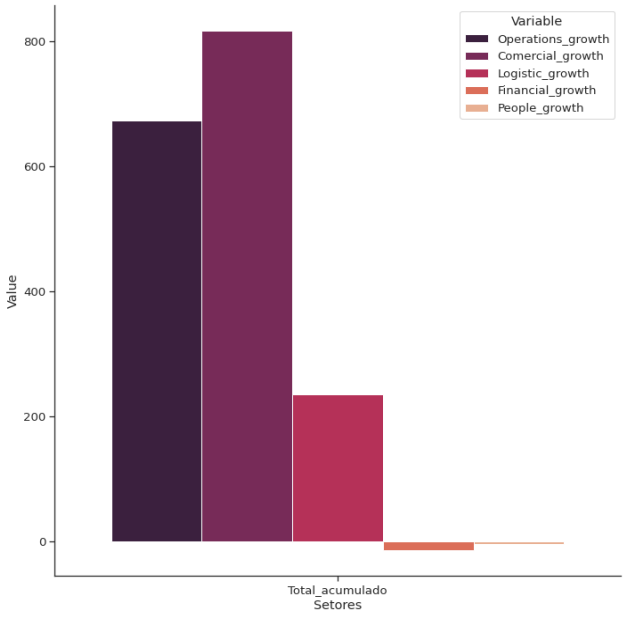


Provavelmente trata-se de uma empresa de vendas. Ela iniciou suas atividades em 2017 com aproximadamente 70 funcionários por setor, distribuídos igualmente. Ao longo dos dois semestres seguintes operaram contratações nos setores que mais importam aos negócios, como operações e comercial (com ~600 funcionários) ou logística (~200 funcionários).

4) Quando observamos a taxa de contratação (e dispensas) ao longo dos semestres, podemos observar que no segundo semestre foi quando mais se contratou, concentrando-se na área de operações e comercial. No semestre seguinte, contratou-se principalmente no setor logístico.



Podemos ter uma visão mais clara quando observamos a taxa acumulada de contratações entre o segundo semestre e o último:



5) Os outros setores mantiveram-se praticamente estáveis no número de funcionários. A área de financeira, porém, é a mais instável entre os setores; decresceu o número total de funcionários em todos os períodos, mas houve grande circulação de funcionários – com um total de 84, mas apenas 68 no último período, enquanto haviam 77 no primeiro semestre.

Período	Operações	Comercial	Logística	Financeiro	Pessoas
2º 2017	79	63	69	79	63
1º 2018	407	364	73	77	67
2º 2018	615	583	236	77	66
1º 2019	611	577	231	68	60
Total_CPFs	*630*	*591*	*240*	*84*	*72*

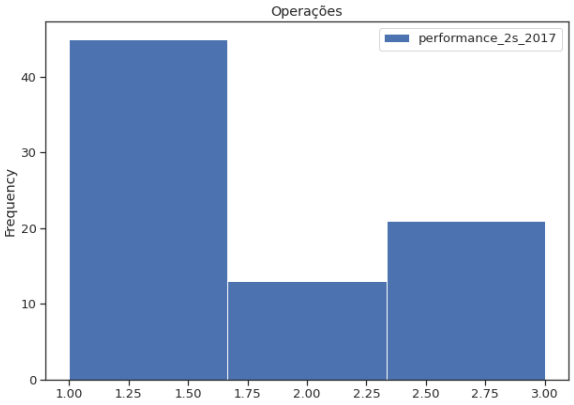
PARTE 2: Performance por Área

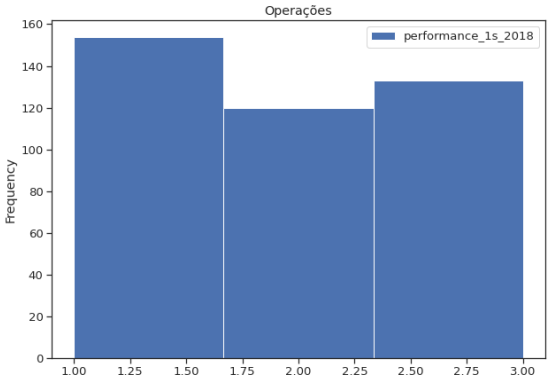
Já sabemos que o negócio da empresa gira em torno de vendas e comércio, concentrando a maior parte dos recursos humanos nesses setores, assim como a logística necessária para a engenharia operacional desse processo. Por outro lado, também sabemos que algumas áreas estão com dificuldade de reter seus funcionários. A partir da análise da performance dos funcionários podemos saber como aproveitá-los melhor de forma que deixe-os satisfeitos, assim como se traduza em resultados para a companhia.

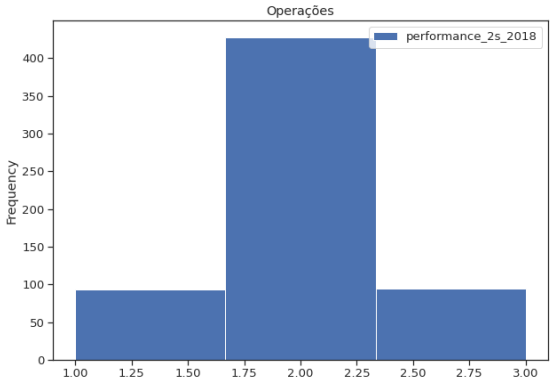
A performance dos usuários foi classificada de 1 a 3. Com isso podemos identificar os funcionários que obtiveram resultado insatisfatório persistente ao longo do tempo e observar em seu teste psicométrico se suas habilidades estão de acordo com a área ou se é melhor transferir para outro setor. Às vezes, suas melhores habilidades estão subutilizadas e podem ser aproveitadas melhor em outro setor.

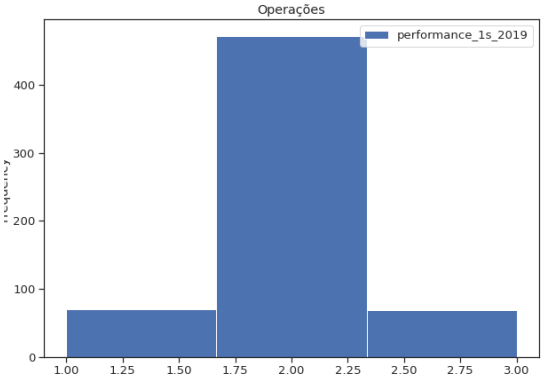
Vamos primeiro observar a distribuição da performance dos usuários por diferentes setores (operações, comercial, logística, financeiro, pessoas). É previsível que se tudo estiver funcionando bem, encontremos uma distribuição normal das performances de grau 1 a 3.

a) Operações:





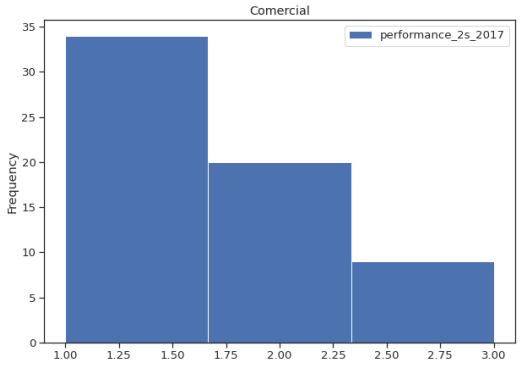


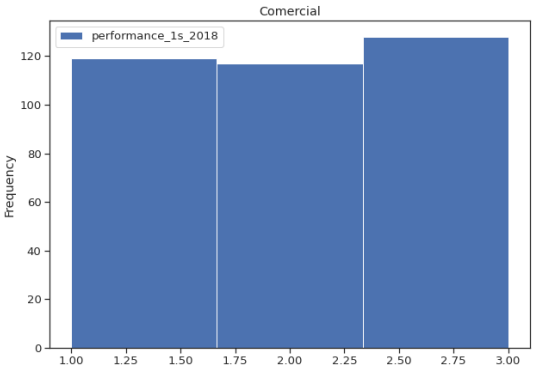


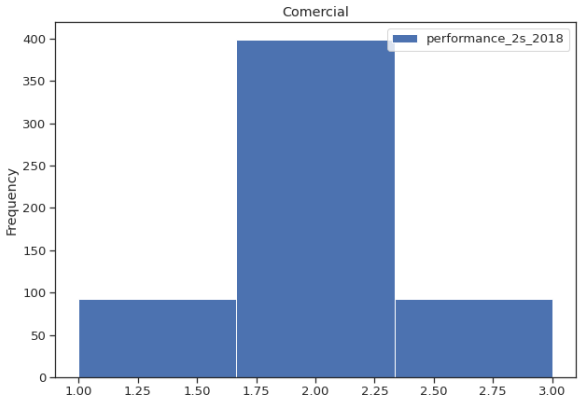
Observando 4 semestres, percebemos que há um período de estabilização progressiva, de resultados ruins e generalizados no primeiro semestre, até algo regular. É nos dois últimos semestres que percebemos uma distribuição normal. É natural que a maior quantidade de trabalhadores esteja no nível 2, já que essa é a média que tomamos como padrão para avaliar as bordas. Porém, percebemos que, apesar do pequeno aumento do quadro de funcionários, no último semestre a média aumentou e a proporção de funcionários com performance ótima (3) diminuiu na mesma medida que os de performance ruim (1).

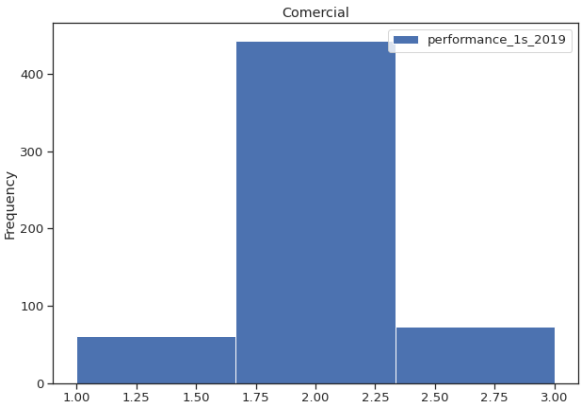
Poderíamos, então, observar quais são os funcionários que tiveram performance ruim e alguns que tiveram performance média (dentro de um threshold aceitável), quais possuem habilidades que possam ser melhor aproveitadas em outras áreas.

b) Comercial:





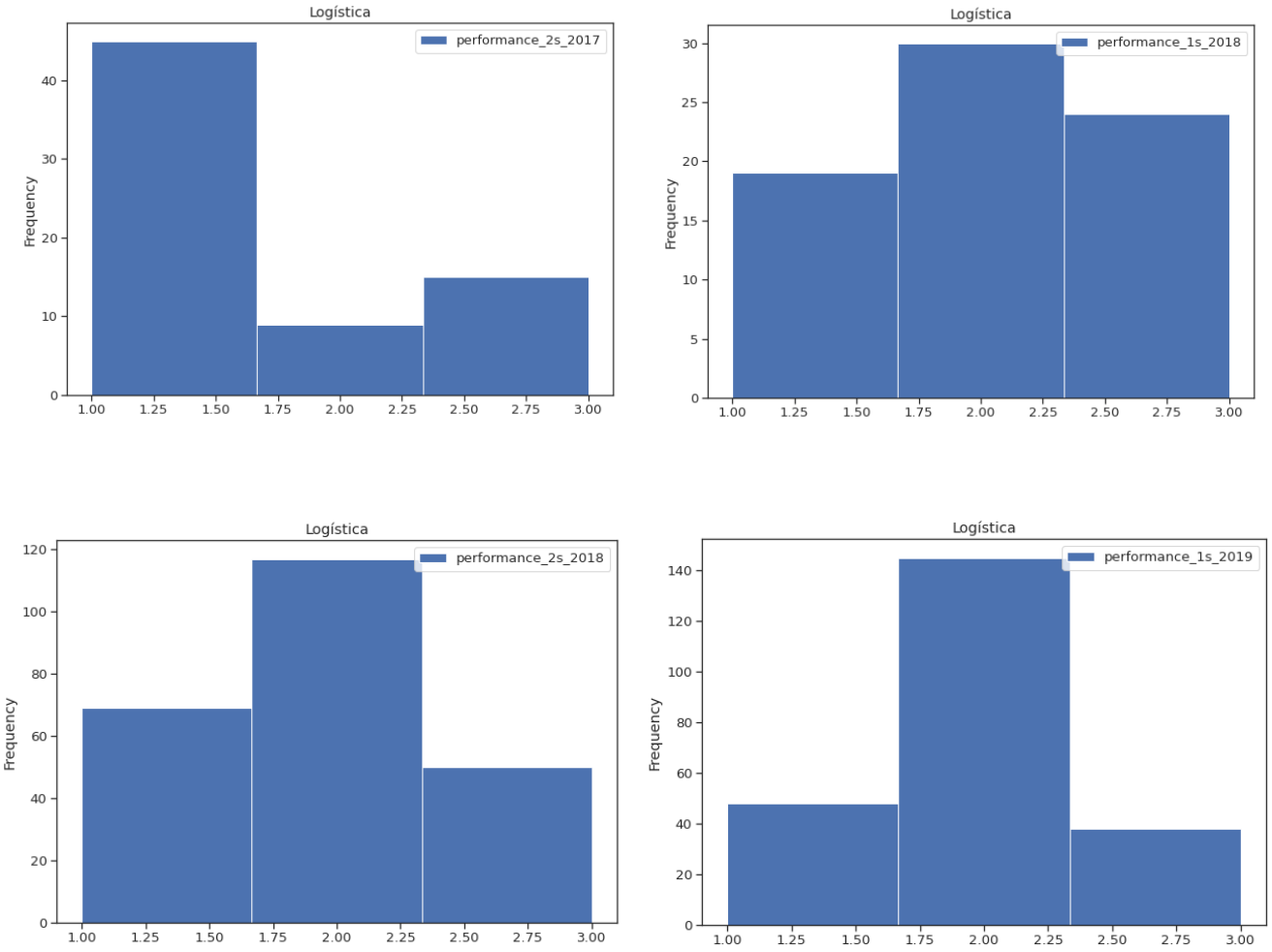




No setor comercial seguiu-se praticamente a mesma lógica do setor de operações. Provavelmente após um período de adaptação e organização do setor, os muitos funcionários que foram contratados no setor pelo período inicial foram se adequando; Ao fim, sobra uma distribuição normal da performance dos funcionários do setor, com um pouco mais de prevalência de performances ótimas sobre performances ruins.

Assim como em Operações, poderíamos observar quais são os funcionários que tiveram performance ruim e alguns que tiveram performance média (dentro de um threshold aceitável), quais possuem habilidades que possam ser melhor aproveitadas em outras áreas.

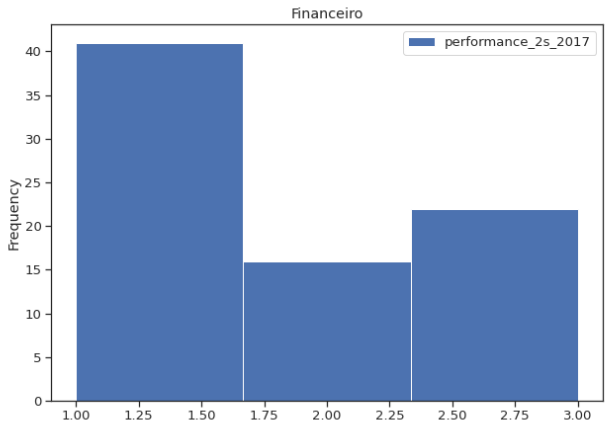
c) Logística:

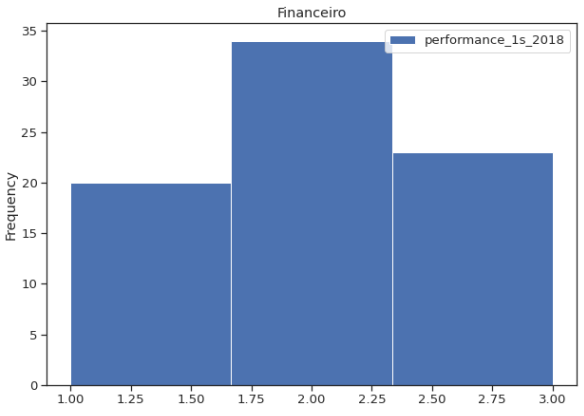


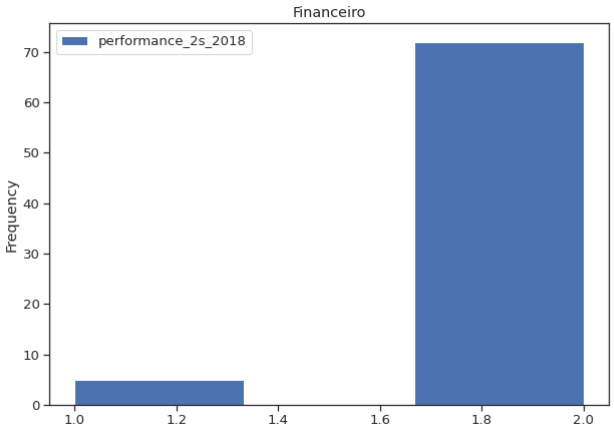
Em logística, nossa terceira maior área de empenho de recursos humanos, seguiu-se também a lógica de progressiva regularização das performances ao longo dos semestres. Porém, é evidente que há uma ineficiência em recursos humanos. Apesar de existir uma boa quantidade de trabalhadores com performance alta (3), é a quantidade de trabalhadores com performance baixa (1) que preocupa. Provavelmente as notas estão adequadas, já que é uma área que permite uma avaliação mais objetiva dos resultados.

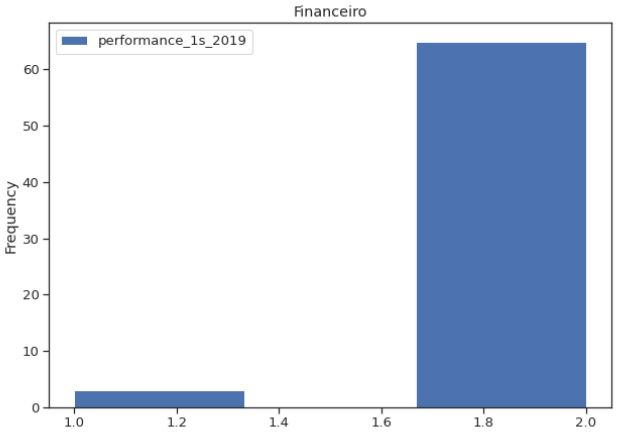
Esse é o setor de maior relevância que temos atualmente e que exige intervenção. Além de ser uma área estratégica com bastante recursos humanos empenhados, a produtividade está mal distribuída. Podemos observar por volta de 50 funcionários que possam ser realocados a partir de suas habilidades que estão subutilizadas.

d) Financeiro:





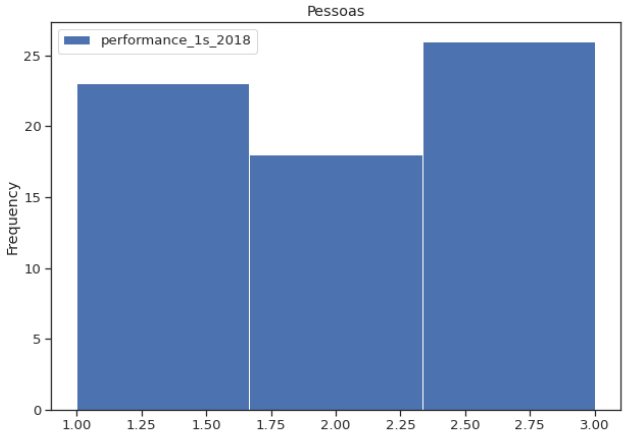


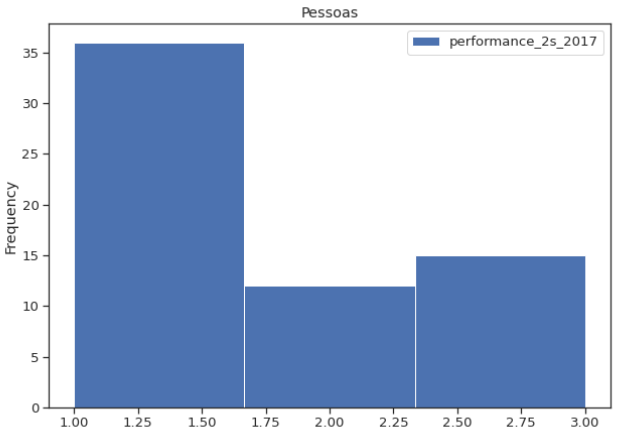


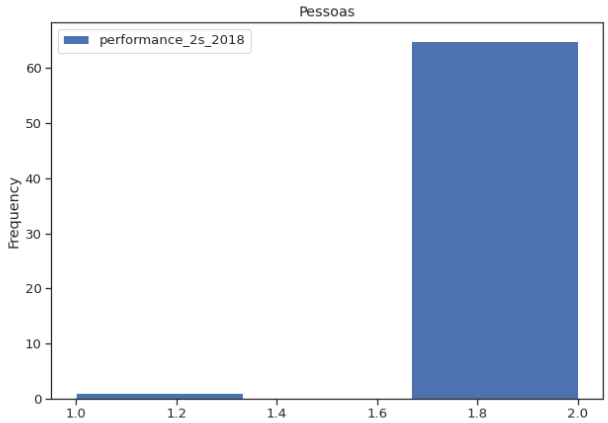
O financeiro faz parte das duas áreas minoritárias da empresa, o que é comum em áreas administrativas. Também, a avaliação de desempenho pode ser mais difícil de avaliar, pelos critérios de resultados serem mais intersubjetivos ou também sensível a externalidades que estão fora do controle da empresa. Também por isso, pode ser uma área onde possui maior circulação.

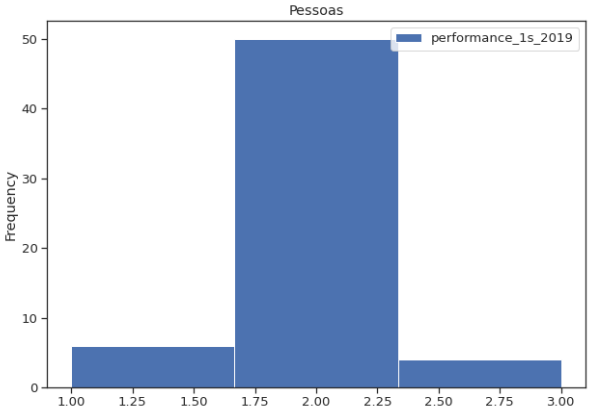
O nosso caso preocupa. Ela nunca estabilizou. Não possui nenhum funcionário com performance ótima (3) como exemplo do que seguir para área. Todos tiveram performance razoável ou baixa. É uma área que está em baixa histórica na permanência dos funcionários, a área já possui poucos funcionários, mas alguns poucos podem ser realocados para outros setores. *O principal, porém, é que funcionários de outras áreas com habilidades excepcionais possam ser trazidos de outros setores em que estão sendo subutilizados, para estabilizar a área financeira*.

e) Pessoas:









É o menor setor da empresa. Normalmente é mais difícil de avaliar resultados, mas cada vez temos critérios mais objetivos para o setor. Ele conseguiu estabilização ao longo dos semestres, mas ainda possui uma quantidade razoável de performance ruim comparada à quantidade de performances excelentes. Também permite observar a realocação desses funcionários, mas já possui poucas pessoas e bastante circulação.

PARTE 3: Habilidades e Realocação

1) Habilidades

Os funcionários da companhia responderam questionários psicométricos em que podemos avaliar suas diversas skills como Raciocínio, Social, Motivacional e nível Cultural/Educacional. Através desses dados, podemos criar scores que permitem **(a)** identificar de maneira objetiva quais são as habilidades mais requeridas por área; **(b)** identificar quais são as possíveis realocações, cruzando com as performances dos trabalhadores e suas respectivas áreas atuais.

Observando os dados, a melhor maneira que encontramos foi buscar a mediana de cada uma das *skills* por área, de forma a identificar qual é o destaque que é dado por um trabalhador razoável para a área que difere das outras áreas. *Sabemos também que a nossa área mais urgente para observar é o setor* **financeiro**, pela má performance que não foi resolvida ao longo do tempo e pela pouca permanência dos funcionários.

Área	Potencial_Bruto	Raciocínio	Social	Motivacional	Cultura
1 *operações*	51.24	49.56	50.82	51.56	44.62
2 *comercial*	50.47	55.79	47.53	51.56	42.03
3 *logística*	54.59	56.15	50.82	54.15	47.65
4 *financeiro*	50.43	52.67	54.12	51.56	50.52
5 *pessoas*	49.89	53.39	54.11	51.56	51.11

Assim, observamos que o maior nível cultural é o de **pessoas** seguido do **financeiro** (urgente), exigindo pessoas com maior nível educacional de outros setores para serem realocados. Na verdade, o funcionário médio mais semelhante nas diversas *skills* com o **financeiro** é do setor de **pessoas**. Os trabalhos provavelmente possuem perfil semelhante (mais administrativo), possuem quantidade de funcionários parecidas, habilidades, nível educacional. Um ponto de atenção é que o setor de pessoas também poderia melhorar com alguns ajustes; outro ponto é que, já que ambos os setores possuem poucos funcionários, é mais arriscado realocar e transferir responsabilidades.

O setor **comercial** e **logístico** também parecem ser intercambiáveis, já que seu perfil médio é semelhante, principalmente exigindo boa capacidade de *Raciocínio*, ou seja, um perfil mais técnico e operacional, por vezes mais quantitativo. No caso de **operações**, parece ser mais intercambiável com **logística**, já que parece exigir maior habilidade *Social*. De qualquer forma, são setores intercambiáveis que poderão ter realocação, principalmente após realocação para outros serviços urgentes – mas esses setores parecem já estar desempenhando de forma ponderada.

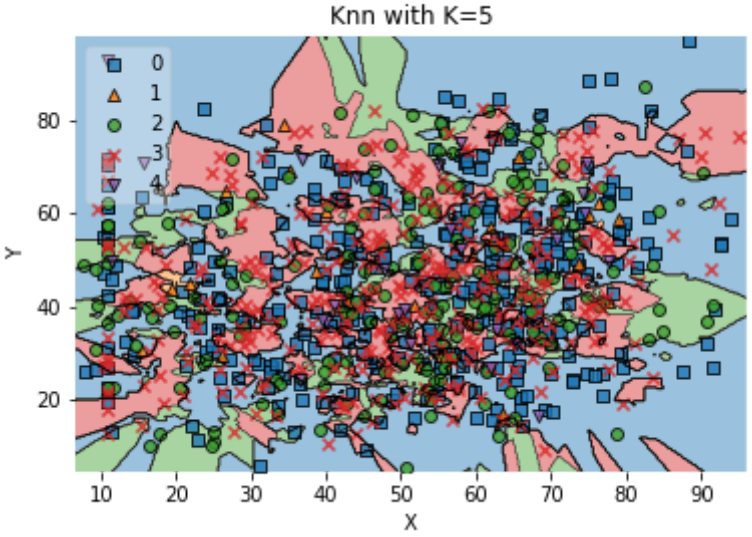
Devemos, porém, fazer uma ***observação importante****.* Os dados não são confiáveis e podem levar a conclusões aleatórias ou enviesadas. Existe uma variância grande entre os valores das colunas e, para piorar, muitas das colunas importantes estão praticamente vazias

(chegando a **89%**[1]) – normalmente, apenas a coluna *Raciocínio* e *Cultura* estão mais completas. Mas, a completude varia muito entre cada área da empresa, o que atrapalha a análise. É ainda pior quando a quantidade de dados varia entre setores que são de diferentes tamanhos e alguns deles são bastante pequenos. Por fim, não foi possível validar se nossas colunas foram calculadas da melhor maneira a partir dos dados brutos e não foi possível derivar thresholds confiáveis para determinar de maneira objetiva quais são os valores que definem as *skills* mais importantes para cada setor.

2) *Auto-realocação: vizinhos mais próximos (KNN)*

Tentamos criar um sistema de recomendação simples para automatizar a realocação usando um modelo de KNN ('*k vizinhos mais próximos*'). Porém, como dissemos, os dados não parecem estar adequados para isso. Decidimos pegar a menor amostra das nossas tabelas que contivessem o máximo de informações possíveis para melhorar o modelo, mas não foi suficiente.

Conseguimos verificar o problema nos dados justamente quando produzimos o nosso modelo e tentamos visualizá-lo com nossos vetores (funcionários + habilidades) em um espaço bidimensional. Podemos observar que os scores de habilidades dos nossos funcionários são todos bastante próximos ou quase aleatórios, não traduzindo em uma boa segregação para formar clusters e, portanto, dificultando automatizar a identificação da melhor área para cada funcionários (podemos observar que não temos boundaries bem definidos nos nossos dados).



Nosso objetivo inicial era utilizar o modelo de KNN de forma que pudéssemos encontrar os funcionários que estivessem na fronteira (*decision boundaries*) com outros clusters, próximo de colegas semelhantes. O algoritmo faz a predição de cada ponto pela semelhança com os outros pontos (vizinhos mais próximos) usando cálculo de distância

[1] Para consultar essas observações e outras, abrir os Notebooks fornecidos em anexo. Infelizmente, não tivemos a oportunidade de refatorar e modularizar o código para melhor compreensão e reprodutibilidade, mas é possível encontrar algumas explorações e insights nessa POC (*Proof-of-Concept*). Optamos aqui, então, por aproveitar os dados que estão mais trabalhados e disponíveis, como Cultura e Raciocínio, de forma a não perder funcionários para análise de realocação. Também, ambas as variáveis são importantes para realocação ao nosso setor mais urgente: o financeiro.

(usamos similaridade de cosseno). Para os que estão na fronteira, o algoritmo opta pelo desempate (não existe ponto de dado sem classificação). Quando pedimos para prever os nossos dados depois do modelo treinado, como ele usa cálculo de distância para predição, às vezes ele pode desempatar sugerindo uma outra classificação que não é a original do ponto de dado (no nosso caso isso não seria exatamente um erro).

Encontramos então 4 sugestões do modelo para realocação para o setor financeiro (nossa área urgente). Seus scores são:

Raciocínio	Cultura	Motivacional	Potencial Bruto	Social	Área	Relocação
42.74	48.54	51.56	51.23	60.7	Comercial	*Financeiro*
31.52	72.07	77.53	45.17	37.65	Operações	*Financeiro*
47.06	53.52	56.75	42.73	27.78	Operações	*Financeiro*
22.96	52.73	41.17	26.89	21.19	Comercial	*Financeiro*

Como podemos perceber, não são scores adequados para melhorar a performance de um setor que está com problemas. Também, não acreditamos que podemos confiar nos dados para um bom modelo de recomendação. Entendemos que a melhor abordagem não seria através do aprendizado de máquina, mas através de uma tarefa analítica de interpretação dos dados.

3) RESULTADOS: Realocações

(a) Funcionários Experientes

Nossa primeira proposta é buscar os funcionários da empresa que tiveram *performance ruim* (**1**) no último semestre, mas que possuem nível educacional (*Cultura*) e *Raciocínio* maior que **50 pontos**. Acreditamos que esses são candidatos que possuem capacidade de entregar mais - e estão mais próximos das exigências das nossas áreas mais urgentes e com poucos funcionários (financeiro e pessoas).

Para agilizar o trabalho de encontrar os melhores candidatos entre a nossa lista, ordenamos primeiro por *Cultura*, depois por *Raciocínio* e depois por *Área*. Sabemos como a dedicação para a formação educacional exige bastante tempo, recursos financeiros e esforços. Isso significa que o funcionário provavelmente está insatisfeito com sua situação, já que sua avaliação de performance foi baixa; considerando quanto dedicou-se para a sua formação, provavelmente estaria mais disposto para realocação, sendo vantajoso para a empresa e funcionário. Além disso, caso sua performance melhore, é possível que sua experiência leve a assumir mais responsabilidades no novo setor e que surjam novas *lideranças*, principalmente nos setores mais necessários.

*Apenas com isso conseguimos retornar **35** possíveis candidatos*. Esse já é um número razoável para entrarmos em contato e avaliar a possibilidade de realocação. O ideal seria buscar funcionários das áreas com mais recursos humanos, como Operações, Comercial e Logística, para não fragilizar áreas menores como Financeiro e Pessoas. Segue a lista ordenada por importância para buscar os candidatos:

	CPF	Área	performance_1s_2019	Potencial Bruto	Raciocínio	Social	Motivacional	Cultura pontuação
0	583.106.279-12	Comercial	1.000000	nan	77.200000	nan	nan	86.990000
1	743.208.165-44	Logística	1.000000	61.270000	51.600000	60.700000	77.530000	81.320000
2	531.296.470-15	Operações	1.000000	56.340000	73.530000	47.530000	41.170000	76.330000
3	970.652.418-58	Comercial	1.000000	59.270000	64.630000	47.530000	67.140000	76.310000
4	726.540.193-14	Logística	1.000000	nan	77.810000	nan	nan	73.850000
5	389.521.764-64	Operações	1.000000	44.800000	74.780000	9.440000	46.360000	73.610000
6	529.438.706-92	Logística	1.000000	47.170000	56.740000	44.240000	35.970000	73.090000
7	735.680.219-86	Pessoas	1.000000	59.670000	54.830000	67.280000	56.750000	70.470000
8	168.457.923-64	Operações	1.000000	nan	72.520000	nan	nan	69.960000
9	026.897.543-47	Operações	1.000000	43.200000	52.200000	34.360000	41.170000	69.430000
10	649.781.305-57	Operações	1.000000	nan	57.380000	nan	nan	67.850000
11	037.146.985-66	Operações	1.000000	60.490000	66.990000	70.580000	35.970000	64.060000
12	387.204.915-14	Logística	1.000000	71.430000	71.530000	57.410000	90.890000	63.260000
13	764.013.825-62	Comercial	1.000000	nan	65.150000	nan	nan	62.410000
14	652.134.809-42	Pessoas	1.000000	nan	56.150000	nan	nan	62.030000
15	259.631.047-34	Logística	1.000000	64.480000	53.140000	60.700000	87.920000	60.890000
16	948.156.073-21	Financeiro	1.000000	nan	65.220000	nan	nan	60.780000
17	876.201.349-13	Operações	1.000000	nan	72.760000	nan	nan	58.940000
18	280.143.695-13	Operações	1.000000	nan	55.850000	nan	nan	58.870000
19	682.097.513-95	Operações	1.000000	nan	55.910000	nan	nan	58.500000
20	465.081.392-15	Operações	1.000000	nan	78.650000	nan	nan	58.080000
21	785.039.461-57	Comercial	1.000000	nan	81.580000	87.040000	nan	57.310000
22	658.073.192-31	Comercial	1.000000	nan	78.800000	nan	nan	56.250000
23	254.896.307-11	Operações	1.000000	61.200000	75.950000	47.530000	56.750000	56.200000
24	704.986.532-18	Comercial	1.000000	nan	82.180000	nan	nan	55.870000
25	518.394.072-12	Operações	1.000000	nan	70.690000	nan	nan	55.370000
26	094.518.263-51	Operações	1.000000	nan	52.230000	83.740000	nan	55.250000
27	945.380.167-48	Operações	1.000000	nan	86.130000	nan	nan	54.730000
28	190.624.783-87	Operações	1.000000	52.860000	61.590000	47.530000	46.360000	54.300000
29	824.601.753-53	Logística	1.000000	58.310000	64.410000	67.280000	35.970000	53.790000
30	056.831.479-48	Comercial	1.000000	nan	63.450000	nan	nan	53.010000
31	290.574.861-31	Logística	1.000000	nan	59.210000	nan	nan	52.940000
32	508.194.736-93	Pessoas	1.000000	nan	52.990000	nan	nan	52.920000
33	340.759.628-65	Comercial	1.000000	nan	74.860000	nan	nan	52.540000
34	261.493.075-14	Logística	1.000000	nan	69.760000	nan	nan	51.790000

(b) Novos Funcionários

Optamos também por investigar os novos funcionários que ainda não tiveram sua performance avaliada. Acreditamos que podemos encontrar promessas que podem se interessar em um programa de aceleração de carreira, dada suas habilidades e formação. Esses profissionais com grande potencial podem compor as equipes que serão reestruturadas com funcionários mais experientes da empresa, nos setores que exigem isso, como o Financeiro.

Uma vez que possuímos um perfil completo de alguns, identificamos aqueles que possuem um potencial bruto como uma promessa na empresa. Encontramos 13 candidatos que são nossas promessas e poderão ser treinados nas novas equipes:

	CPF	Área	Potencial Bruto	Raciocínio	Social	Motivacional	Cultura pontuação
0	392.418.576-17	Comercial	82.710000	94.050000	77.160000	72.340000	58.890000
1	143.956.027-71	Comercial	75.150000	82.700000	57.410000	87.920000	53.540000
2	178.259.046-31	Comercial	74.130000	70.680000	63.990000	93.860000	60.620000
3	189.560.273-41	Logística	73.570000	71.570000	80.450000	67.140000	70.310000
4	607.814.395-66	Comercial	72.340000	70.630000	70.580000	77.530000	53.680000
5	645.387.902-65	Comercial	71.150000	73.800000	67.280000	72.340000	59.310000
6	521.498.307-79	Comercial	71.050000	67.790000	73.870000	72.340000	51.790000
7	654.712.839-37	Operações	70.790000	58.400000	70.580000	90.890000	58.230000
8	643.805.291-42	Comercial	70.120000	77.710000	67.280000	61.950000	65.170000
9	643.902.578-38	Comercial	65.470000	83.730000	50.820000	56.750000	51.520000
10	193.782.045-97	Operações	60.720000	59.230000	54.120000	72.340000	73.590000
11	410.873.956-66	Comercial	58.880000	51.010000	50.820000	82.730000	62.520000
12	849.035.726-92	Comercial	57.480000	57.270000	50.820000	67.140000	51.730000
13	809.713.654-66	Operações	54.430000	56.490000	54.120000	51.560000	61.600000

Também separamos outras 6 promessas que não possuímos seus perfis completos, mas que possuem um destaque raro em habilidades essenciais como Raciocínio e Cultura, com pontuação sempre acima de 70% (muito acima da média):

	CPF	Área	Raciocínio	Cultura pontuação
0	481.967.352-19	Operações	70.870000	77.680000
1	253.018.467-44	Comercial	89.330000	76.730000
2	764.812.950-76	Operações	73.880000	76.610000
3	642.713.509-15	Comercial	74.960000	74.920000
4	237.861.950-21	Comercial	77.540000	74.410000
5	067.234.859-47	Comercial	83.800000	73.840000

CONCLUSÃO: Recomendações para Trabalhos Futuros

*Nossa primeira recomendação é a **recoleta** dos dados e fortalecer uma cultura centrada nos dados*. É importante entender o *porquê* dos trabalhadores não terminarem as *surveys* e colunas importantes estarem praticamente incompletas. Talvez outra abordagem seja necessária para garantir. Também é necessário que as *surveys* sejam criadas com auxílio dos analistas e cientistas de dados que vão usá-las, para identificar o quê e de que maneira é melhor ser coletado[2]. Além disso, essa aproximação permite um melhor entendimento e confiança nos dados quando for analisado, de forma que possam ser sempre validados; com isso podemos ter material para *verificação e falseabilização de hipóteses* a partir da capacidade de reproduzir experimentos (reprodutibilidade). Outro ponto importante é criar uma ***documentação*** adequada para permitir padronização na interpretação dos dados e continuidade das análises de forma que seja menos dependente das pessoas. Também é necessário ***padronizar os dados*** e o banco de dados para facilitar o trabalho.

Uma vez que tivermos bons dados, podemos avançar em tentativas que ainda não obtiveram sucesso. Podemos, por exemplo, **identificar exatamente quais CPFs são candidatos confiáveis** para realocação e não apenas sugerir de quais áreas eles podem vir.

*Mais, uma vez que tivermos melhores dados, podemos **modelar** o processo de sugestão de realocação e contratações*. Poderíamos, por exemplo, criar modelos supervisionados baseados em algoritmos de classificação hierárquicos como *CatBoost* para, dada às características de um funcionário, sugerir qual é a área adequada para trabalhar. Mais do que isso, poderíamos criar modelos mais complexos que um KNN, ou para outras funções como encontrar os funcionários mais semelhantes que ainda não colaboram juntos. Talvez, criar um *sistema de recomendação*. Poderíamos mesmo estabelecer *séries históricas* ou *forecasting* se houvesse maior regularidade nos dados temporais como no caso dos scores de performance. Se houver caixa de texto para os pesquisados escreverem, também é possível incorporar modelagem textual através do processamento de linguagem natural.

*Por fim, a longo prazo, uma cultura de dados possibilitaria a engenharia de aprendizado de máquina para **automatizar** esses processos de forma recorrente*, através da infra-estrutura de serviços na nuvem (*GCP, AWS, Gitlab*) em que podemos usar ferramentas como *CI/CD, kubeflow, airflow, dataflow, mlflow, weight&bias* etc.

---

[2] Acreditamos, por exemplo, que seria interessante incorporar dados básicos como gênero, cor, orientação sexual, origem, para observar a pluralidade das equipes ou mesmo estimulá-las, colhendo os benefícios da diversidade.

FEEDBACK

Agradeço pela oportunidade de participar desse processo seletivo. Não pude concluí-lo conforme seria ideal. *Gostaria de deixar um **feedback** que seja **construtivo***.

1) Esse é um dos *cases* mais frustrantes que já peguei. Minha intenção é fazer uma crítica construtiva e natural. Isso é uma opinião fruto dos meus sentimentos, pensamentos e observações conforme fui tentando concluí-lo e minha observação vem dessa experiência a partir de um *conhecimento situado*. Espero que a minha experiência, no meio de tantas várias que divergem da minha, possa ajudar em algo.
   a) O formato fornecido em planilha de **excel** (com mais de uma tabela no arquivo) é incômodo e incomum; geralmente seriam arquivos **CSV** ou outros, ou normalmente um banco de dados com acesso via **SQL**;
   b) Vocês não forneceram nem os **dados brutos** (tabela com as respostas para as perguntas) para dar autonomia e tampouco forneceram os **dados bem processados** permitindo concentrar-me em alguma análise ou modelagem específica para uma resposta provável de ser avaliada de forma objetiva. O melhor seria um ou outro e não o meio termo;

2) Pela minha experiência, normalmente é fornecido um case onde os avaliadores já desconfiam onde se pode chegar. Mesmo, normalmente fornecem uma **pergunta mais objetiva** do que 'analise os dados e proponha melhorias na empresa'; por exemplo, uma vez pediram para eu trabalhar um conjunto de dados e criar um modelo preditivo para predição de cancelamento de reserva de hotéis, outra vez para classificação de diagnósticos, outro para redução de custos a partir de dados logísticos bem construídos.

3) Existe uma diferença do trabalho de um **analista de dados** e do **cientista de dados**. Com os dados que vocês forneceram é muito difícil fazer um trabalho de cientista de dados, como é o de **modelagem**. Mesmo para análise dos dados é difícil criar um trabalho interessante em cima da qualidade desses dados.

4) Vocês forneceram tabelas mais ou menos processadas sem **documentação** nenhuma. Isso não condiz com as boas práticas de dados, de programação ou de pesquisa. O correto é sempre ter as colunas documentadas, ainda mais quando já foram processadas; seja se é para um banco de dados (existem descrições para as colunas), seja se é para a coleta de *surveys* seguida de análises, sempre em uma pesquisa científica é fornecido um documento descrevendo a metodologia da coleta e as descrições das colunas. Essa parece ser inclusive a prática na área de psicologia e psicometria, de testes de personalidade, como outros projetos que conduzi e pratiquei a partir desse banco de dados: https://openpsychometrics.org/_rawdata/ .
   a) Assim, perde-se **tempo** demais limpando os dados e investigando para tentar descobrir o que significa cada uma das variáveis e o que significa os valores que estão nelas; de qualquer forma, apesar da experiência poder ajudar, tudo isso continua a ser intuitivo e propenso a erros, além de limitar bastante a capacidade criativa pela necessidade de manter o ceticismo científico – mesmo se fossemos arriscar adivinhar, não se pode muito. Se a intenção é avaliar essa

capacidade, acredito que desperdiça a avaliação das habilidades que geram mais valor e que exigem muito tempo de treinamento/estudo;

5) Não está claro se esses são dados criados artificialmente – é importante saber isso para analisar;

a) Se realmente são **artificiais**, eles não foram criados com cuidado para permitir uma exploração dos dados nas tarefas mais relevantes para a área, gastando esforços demais na limpeza dos dados que estão muito sujos (até os nomes das colunas atrapalham, com espaços, maiúsculas, despadronizadas, estão fora das boas práticas de dados);

b) Os dados não parecem colaborar para modelagens que poderiam ser interessantes: possui um período pequeno para fazer uma série histórica ou *forecasting*, possui poucos CPFs para alguma modelagem;

c) Mesmo com poucos CPFs poderíamos pensar em alguns modelos que contornassem isso (ex: abordagens *bayesianas*), mas as linhas da tabela em sua maioria estão **vazias** nas informações mais relevantes. Inclusive, parece ser insuficiente para fazer um bom *feature engineering* ou *data augmentation*. Quando segregamos os dados por *área* e *performance* para tentar identificar quais habilidades cada área pede, tanto os que tiveram *performance ruim* como os que tiveram *performance boa* quase em totalidade estão ausentes as colunas relevantes (com exceção de *Raciocínio*) para identificar isso de uma forma objetiva ou automática – o candidato à vaga ainda não é especialista sobre o assunto para intuir e mesmo faz parte do trabalho do cientista de dados prevalecer o ceticismo e basear-se acima de tudo nos dados;

d) Mesmo os dados de *performance razoável* (*2*) que naturalmente são maioria, ainda possuem poucas linhas completas; mesmo que já se soubesse que seria mais difícil inferir as colunas de habilidades mais importantes por área, ao analisar, percebeu-se que essas linhas não geram nenhum insight sobre isso e possuem uma distribuição bastante **aleatória**. Ou seja, sequer conseguimos encontrar soluções realmente confiáveis para transição interna de funcionários entre as áreas das empresas a partir de suas habilidades subutilizadas;

e) A completude dos dados está muito **mal distribuída** entre as áreas; são tantos dados faltantes ou a variância é tão grande que é impossível prever e preencher os valores nulos (por exemplo, com a mediana ou com previsão de regressões lineares). Isso torna impossível criar qualquer modelo preditivo confiável;

f) Talvez fosse possível criar modelos não-supervisionados a partir desses dados, mas seria bastante trabalhoso em curto período de tempo e os dados não parecem confiáveis, isso é bastante arriscado para esses modelos que possuem uma tarefa de avaliação mais difícil, assim como as tentativas de afinar o modelo para os dados;

6) Se os dados são **reais** e apenas anonimizados, algumas coisas preocupam.

a) Não parece que os dados foram realmente **anonimizados** com cuidado, mas posso estar enganado, é difícil saber;

b) Os dados, além de ausentes em partes importantes, possuem uma **representação ruim** e, provavelmente, há problema na **coleta**. Talvez seja necessário retornar, reavaliar e reforçar a etapa de coleta de dados junto aos especialistas (imagino que sociólogos, psicólogos etc) que criaram as *surveys*, de forma a ter perguntas que possuam representações relevantes. Também, investigar os motivos das ausências de respostas e incompletude, para buscar formas de superar isso;

c) A tarefa primária para criar uma **cultura de dados** é coletar dados com qualidade. Em seguida é necessário criar uma infraestrutura adequada para receber, armazenar, processar e disponibilizar esses dados. Apenas após isso podemos ter uma análise adequada desses dados, assim como a possibilidade de criar modelos confiáveis. Se isso já está consolidado, é até mesmo possível criar uma infraestrutura que escale e automatize o aprendizado de máquina. Sabemos que é raro possuir todas essas capacidades, mas ela deve ser um norte;

7) Talvez analistas e cientistas de dados bastante experientes (comparados a um jovem cientista de dados) consigam trabalhar melhor os dados disponibilizados e tirar conclusões relevantes, mesmo para criar modelos; mas mesmo para eles, acredito que esses problemas **atrasem** e **atrapalhem** o seu trabalho. Mesmo que consigam, é difícil confiar se as conclusões são boas representações da realidade, pela má qualidade dos dados.

8) De qualquer forma, o **tempo** é pouco para tentar criar algo interessante a partir da qualidade dos dados disponibilizados. É exigido muito trabalho de limpeza. Além disso, a falta de documentação dificulta muito o processo – isso seria uma prática inadequada mesmo em um ambiente real de trabalho, onde há circulação de profissionais diante da dinâmica do mercado de trabalho, são práticas reiteradas pelos desenvolvedores de software e profissionais de dados (como na filosofia *Agile* e metodologia *SCRUM*, por exemplo).

9) Mesmo para um **case** de **processo seletivo**, espera-se que sejam cases que possam ser de alguma forma **divertidos**, bons para praticar a criatividade, a abstração – mas nesse caso houveram muitos percalços no caminho (que poderiam ser um pouco toleráveis no cotidiano real do trabalho) que impossibilitam dedicar o pouco tempo em tarefas que realmente agregam valor.