



Manipulação de Dados do Recursos Humanos Com Linguagem Python

16.04.2022

Fausta Joaquim Faustino
Analista de dados

São Paulo, SP

Visão Geral

O papel principal dos recursos humanos é administrar funcionários visando os objetivos da empresa com os colaboradores e, promover a satisfação profissional. Além disso, por meio da análise de dados e com o auxílio de softwares de gerenciamento, os recursos humanos tornaram-se uma peça-chave no desenvolvimento do capital humano na empresa.

Definição do problema

Este projeto visa estudar a relação de diversos atributos de funcionários de todos os setores da empresa e o impacto na promoção de funcionários. A empresa coletou dados de um ano, para servir de banco de dados para responder às seguintes perguntas:

1. - Qual a Correlação Entre os Atributos dos Funcionários?
2. - Qual o Tempo de Serviço da Maioria dos Funcionários?
3. - Qual Avaliação do Ano Anterior Foi Mais Comum?
4. - Qual a Distribuição das Idades dos Funcionários?
5. - Qual o Número de Treinamentos Mais Frequente?
6. - Qual a Proporção dos Funcionários Por Canal de Recrutamento?
7. - Qual a Relação Entre a Promoção e a Avaliação do Ano Anterior?

Conjunto de dados

O dataset fornecido para este projeto contém 54 808 linhas e 11 colunas de dados dos funcionários da empresa, conforme pode ser analisado pelo cabeçalho do dataset RH que é apresentado na Tabela 1, com as presentes variáveis:

1. **id_funcionario**: ID dos funcionários ;
2. **departamento**: Departamento em que cada funcionário está alocado;
3. **região**: Região em que se encontra cada funcionário;
4. **educação**: Nível de educação de cada funcionário;
5. **genero**: Gênero de cada funcionário;
6. **canal_recrutamento**: Canal por onde o funcionário foi recrutado;
7. **numero_treinamentos**: Numero de treinamento que o funcionário realizou durante o tempo em que o funcionário esteve na empresa;
8. **idade**: Idade de cada funcionário;
9. **aval_ano_anterior**: Avaliação de cada funcionário;
10. **tempo_serico**: Tempo que o funcionário trabalha na empresa;
11. **promovido**: promoção do funcionário dentro da empresa.

Tabela 1: Conjunto de dados do RH

	id_funcionario	departamento	regiao	educacao	genero	canal_recrutamento	numero_treinamentos	idade	aval_ano_anterior	tempo_servico	promovido
0	65438	Marketing	Sudeste	Mestrado ou Doutorado	F	Outsourcing	1	35	5.0	8	0
1	65141	Logistica	Nordeste	Superior	M	Outro	1	30	5.0	4	0
2	7513	Marketing	Sudeste	Superior	M	Outsourcing	1	34	3.0	7	0
3	2542	Marketing	Norte	Superior	M	Outro	2	39	1.0	10	0
4	48945	Tecnologia	Nordeste	Superior	M	Outro	1	45	3.0	2	0

Análise exploratória dos dados / Limpeza e transformação dos dados

Na etapa de limpeza e transformação dos dados, optei por verificar a existência de valores ausentes em cada uma das linhas e obter o somatório dos valores ausentes em cada variável. Essa verificação permitiu observar que a maioria das variáveis tem 0 valor ausente e, a variável **educação** e a variável **aval_ano_anterior**, tem uma grande quantidade de valores ausentes. Valores ausentes nas variáveis tornam-se um problema para o nosso conjunto de dados, portanto, esse problema não pode seguir adiante.

Para resolver o problema de valor ausente optei inicialmente em realizar o agrupamento de cada uma das variáveis que cometem os valores ausentes. A primeira variável a ser agrupada é a variável **educação**, que está apresentada na tabela 2 conforme mostrando o valor total de registo para cada valor da variável.

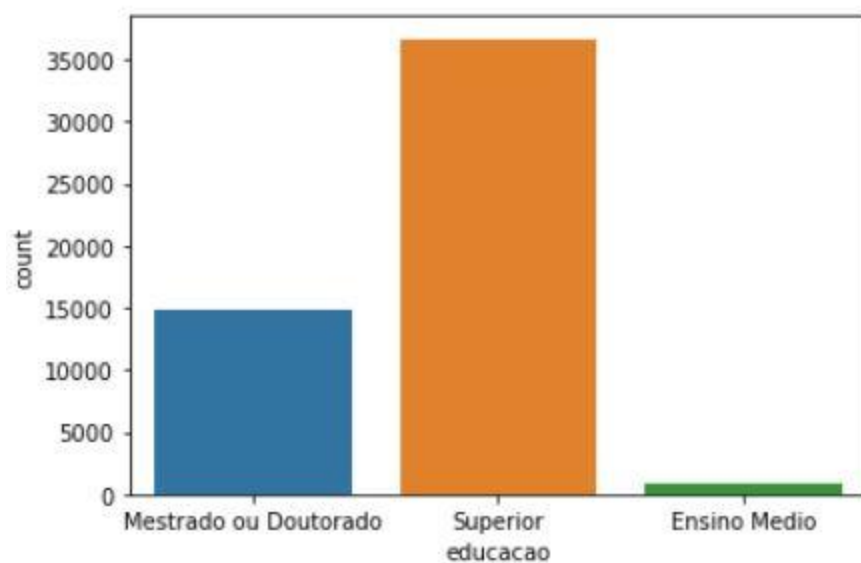
Tabela 2. Total de registo para cada valor da variável educação sem valores ausentes

	id_funcionario	departamento	regiao	genero	canal_recrutamento	numero_treinamentos	idade	aval_ano_anterior	tempo_servico	promovido
educacao										
Ensino Medio	805	805	805	805	805	805	805	458	805	805
Mestrado ou Doutorado	14925	14925	14925	14925	14925	14925	14925	14798	14925	14925
Superior	36669	36669	36669	36669	36669	36669	36669	33404	36669	36669

A tabela 2, mostra que a empresa tem um total de 805 funcionários com ensino médio, 14 925 funcionários com mestrado e doutorado e 36 669 funcionários com ensino superior, totalizando 52 399 funcionários. Esse valor mostra que o agrupamento realizado não considera o número de valores ausente, isso porque o número total de funcionários é 54 808.

Para melhor compreensão dos valores da tabela 2, um gráfico é gerado para verificar o total de registo para cada nível de educação. O gráfico da Figura 1 mostra o total de registo dos funcionários com ensino médio, superior e mestrado ou doutorado.

Figura 1. Total de registo para o ensino médio, superior e mestrado ou doutorado

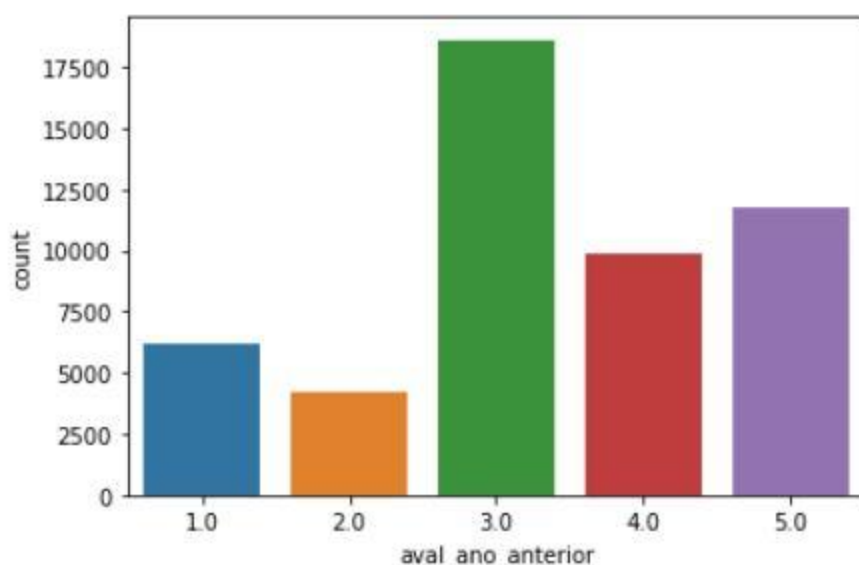


A mesma análise realizada para a variável educação é aplicada na variável aval_ano_anterior que está apresentada na Tabela 3 e no gráfico da Figura 2. A Tabela 3 e a Figura 2 apresentam a nota de avaliação dos funcionários. O gráfico 2 mostra que a maioria dos funcionários se encontra com a nota de avaliação na média.

Tabela 3. Nota de avaliação para todas as variáveis

	id_funcionario	departamento	regiao	educacao	genero	canal_recrutamento	numero_treinamentos	idade	tempo_servico	promovido
aval_ano_anterior										
1.0	6223	6223	6223	5887	6223	6223	6223	6223	6223	6223
2.0	4225	4225	4225	4035	4225	4225	4225	4225	4225	4225
3.0	18618	18618	18618	17867	18618	18618	18618	18618	18618	18618
4.0	9877	9877	9877	9509	9877	9877	9877	9877	9877	9877
5.0	11741	11741	11741	11362	11741	11741	11741	11741	11741	11741

Figura 2. Nota de avaliação dos funcionários.



Após a exploração das variáveis que têm valores ausentes optei por escolher uma técnica para tratar esses valores. A técnica escolhida para tratar valores ausentes é a *imputação* que é uma técnica que irá preencher os valores ausentes da variável educação e a variável val-ano-anterior. Para preencher os valores ausentes da variável educação utilizei a medida de posição moda, que representa o valor que aparece com a maior frequência.

O cálculo da moda é realizado com os dados da tabela 2, onde é desconsiderado os valores ausentes. Já para a variável val-ano-anterior utilizei a mesma técnica de imputação, com a medida de posição mediana. Conclui que a utilização da técnica de imputação permitiu manter o volume de dados original, mas sem ter a presença de valores ausentes.

A segunda etapa de exploração dos dados é realizada na variável promovido, onde é indicada a quantidade de funcionários que foram promovidos. Para a variável promovido optei por realizar o agrupamento para obter a contagem dos funcionários promovidos conforme apresentado na Tabela 4.

Tabela 4. Funcionários promovidos

	id_funcionario	departamento	regiao	educacao	genero	canal_recrutamento	numero_treinamentos	idade	aval_ano_anterior	tempo_servico
promovido										
0	50140	50140	50140	50140	50140	50140	50140	50140	50140	50140
1	4668	4668	4668	4668	4668	4668	4668	4668	4668	4668

Na variável promovido existem duas classes, zero que indica a classe negativa para os funcionários que não foram promovidos, contendo 50 140 pessoas e, a classe um que é a classe positiva com 4 668 pessoas que foram promovidas. Para melhor a compreensão das classes de valores é gerado um gráfico que permite observar a diferença entre as variáveis. O gráfico da Figura 3 mostra que no dataset tem muito mais pessoas sem promoção em comparação com o número de pessoas que foram promovidas.

Se os dados da variável promovido fossem utilizados para treinar o modelo de aprendizado de máquina, o modelo teria muito mais exemplos da classe negativa em comparação com a classe positiva. Portanto, os dados não podem ser utilizados dessa forma porque estaria comprometendo a performance do modelo de aprendizado de máquina. Com essa análise deparei-me com um problema de desbalanceamento de classes. Para resolver o problema de desbalanceamento das classes vou utilizar a técnica *Upsample* que vai permitir criar registros sintéticos de amostragem dos dados para que ao final a classe minoritária fique com o mesmo volume de dados da classe majoritária. O resultado da técnica *Upsample* é apresentado no gráfico da Figura 4, onde é mostrado o balanceamento das classes.

Figura 3. classe positiva e negativa

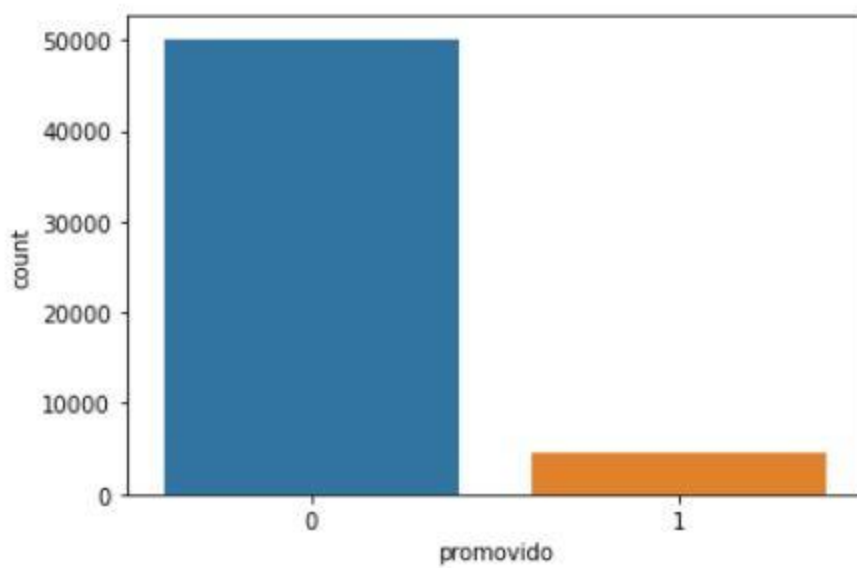
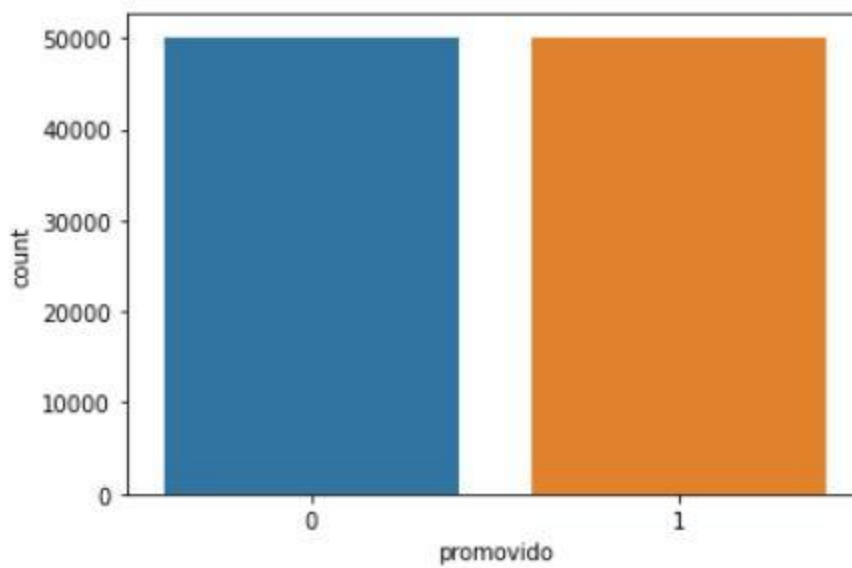


Figura 4. Balanceamento das classes



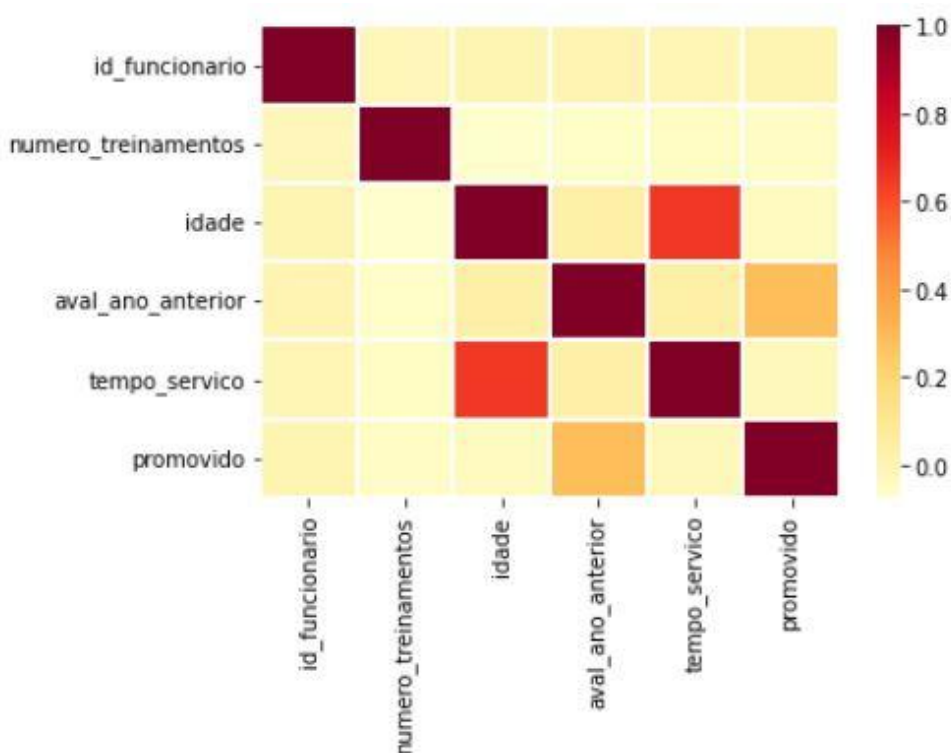
Apresentação de resultados

Nesta etapa serão apresentados os resultados das sete perguntas levantadas pelo RH.

Qual a Correlação Entre os Atributos dos Funcionários?

Para responder a essa pergunta foi criado uma matriz de correlação usando a função *heatmap* do pacote *seaborn* para gerar gráfico estatístico apresentado na Figura 5.

Figura 5. Gráfico de correlação entre os atributos dos funcionários

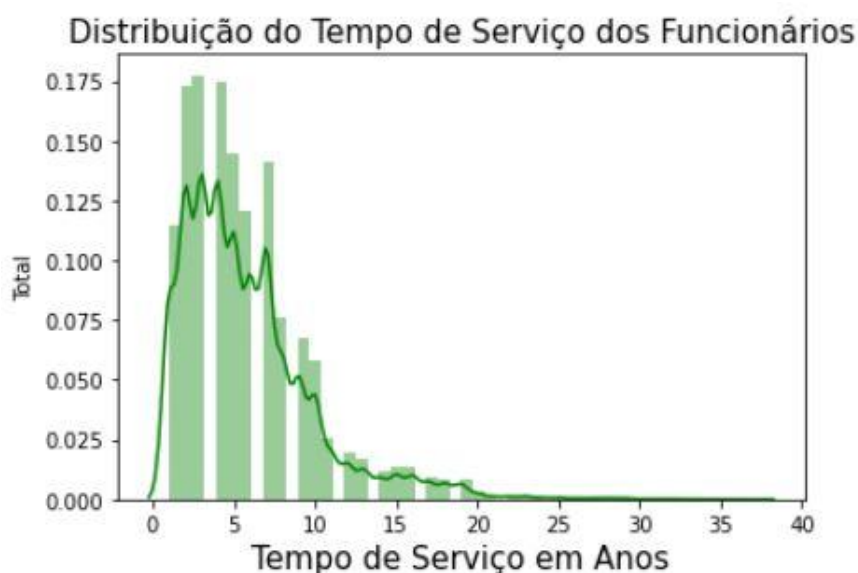


Na figura 5 observa-se que tem uma legenda do lado direito, que vai de 0 a 1 mostrando que a correlação é positiva. Uma correlação positiva indica que quando uma variável aumenta a outra também aumenta e que o valor próximo de zero indica que não há correlação. Acompanhando o mapa é visto que na diagonal mais escura tem valores de correlação igual a 1, indicando, uma correlação de uma variável com ela mesma. Na caixinha vermelha onde tem idade e tempo de serviço é mostrado a correlação de duas variáveis que estão perto de 0,6 e 0,7. Esse valor de correlação indica que quanto maior a idade maior o tempo de serviço, isso porque normalmente as pessoas vão passando mais tempo na empresa.

Qual o Tempo de Serviço da Maioria dos Funcionários?

O tempo de serviço da Maioria dos funcionários é uma pergunta que pode ser respondida através do gráfico de distribuição da variável tempo de serviço. O gráfico de distribuição foi criado através da função *distplot*, para responder a essa pergunta, conforme é apresentado na Figura 6 .

Figura 6. Gráfico de distribuição do tempo de serviço dos funcionários

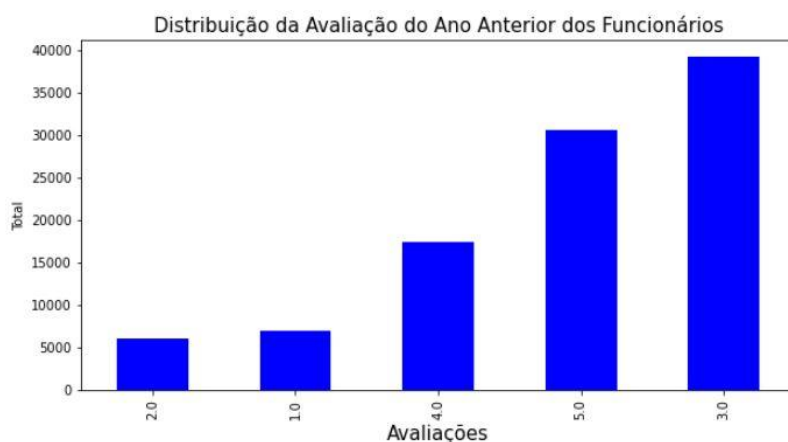


Na Figura 6 o pico da imagem mostra que o tempo de serviço da maioria dos funcionários é de 3 a 5, mostrando que à medida que vai aumentando o tempo de serviço diminui a quantidade de funcionários

Qual Avaliação do Ano Anterior Foi Mais Comum?

Para responder essa pergunta criei um gráfico de barras, para poder analisar a distribuição da avaliação do ano anterior dos funcionários. Para esta empresa a maioria dos funcionários teve a avaliação igual a 3 que é a média.

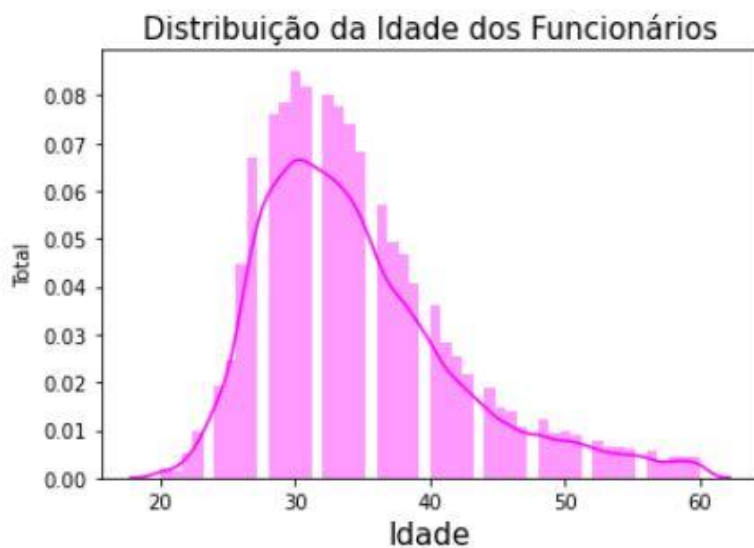
Figura 7. Gráfico de distribuição do ano dos funcionários



Qual a Distribuição das Idades dos Funcionários?

A distribuição das idades dos funcionários está apresentada na Figura 8, onde mostra que a maioria dos funcionários está na faixa dos 30 a 40 anos.

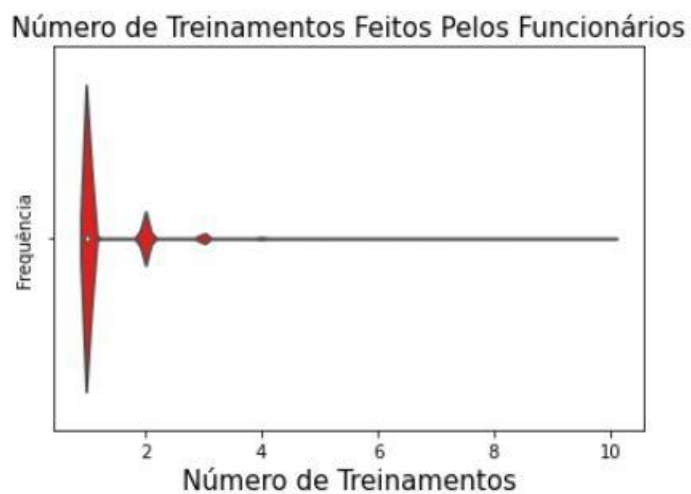
Figura 8. Gráfico de distribuição da idade dos funcionários



Qual o Número de Treinamentos Mais Frequente?

Para resolver essa pergunta criei um gráfico de frequência de distribuição apresentado na Figura 9, onde mostra que a maioria dos funcionários fez apenas um treinamento.

Figura 9. Número de treinamentos feitos pelos funcionários



Qual a Proporção dos Funcionários Por Canal de Recrutamento?

Para responder a essa pergunta criei um gráfico de pizza, que mostra que o maior canal de recrutamento é por outros que atinge 55.22%.

Figura 10. Proporção dos funcionários por canal de recrutamento

