Apresentação de Resultados da Análise de Dados Salariais

Link do GitHub

GitHub

Objetivo

Busca-se a partir do dataset analisado, verificar: (i) quais são os itens que mais influenciam o salário de um empregado e (ii) identificar qual característica influencia mais no salário, escolaridade ou experiência, de forma a direcionar os esforços das personas.

Assim, o objetivo do projeto foi verificar, a partir de uma base de dados salarial, os componentes que mais influenciavam no salário pago a uma pessoa a partir do perfil/características apresentadas, bem como gerar um modelo de regressão linear sobre o caso

Amostra

Uma amostra dos dados iniciais pode ser observada abaixo:

	Age	Gender	Education Level	Job Title	Years of Experience	Salary
0	32	Male	Bachelor's	Software Engineer	5	90,000
1	28	Female	Master's	Data Analyst	3	65,000
2	45	Male	PhD	Senior Manager	15	150,000
3	36	Female	Bachelor's	Sales Associate	7	60,000
4	52	Male	Master's	Director	20	200,000
5	29	Male	Bachelor's	Marketing Analyst	2	55,000
6	42	Female	Master's	Product Manager	12	120,000
7	31	Male	Bachelor's	Sales Manager	4	80,000
8	26	Female	Bachelor's	Marketing Coordinator	1	45,000
9	38	Male	PhD	Senior Scientist	10	110,000

Análise Exploratória

Principais Características Descritivas

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 377 entries, 0 to 376
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Age                  375 non-null    float64
 1   Gender               375 non-null    object 
 2   Education Level      375 non-null    object 
 3   Job Title            375 non-null    object 
 4   Years of Experience  375 non-null    float64
 5   Salary               375 non-null    float64
dtypes: float64(3), object(3)
memory usage: 17.8+ KB
```

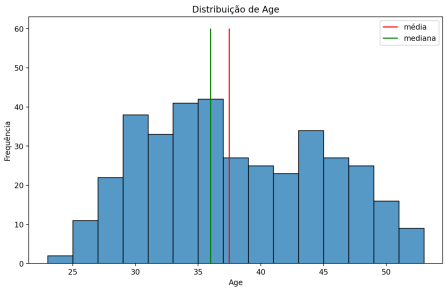
Conforme registrado acima, observou-se dois registros nulos que foram excluídos do dataset.

Análise da Distribuição das Principais Variáveis

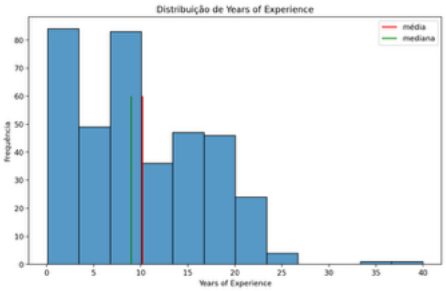
Análise da Distribuição das Principais Variáveis

A análise foi realizada de forma completa para as outras variáveis, sendo verificado inclusive o equilíbrio da base em termos de número de registos para cada categoria. Todavia, Age e Years of Experience se mostraram como as que necessitavam de maior atenção no processo, por isso o destaque.

Age Distribution:



Years of Experience Distribution:

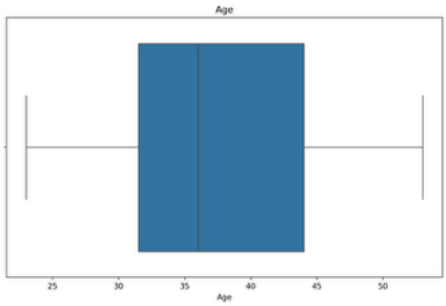


Durante a análise, pôde-se observar que, tanto na feature Age quanto na feature Years of Experience, os valores da média e mediana são próximos, com uma diferença percentual de 4,10% e 13,08% respectivamente. Outro ponto a ser observado é que nos 2 casos a média é maior que a mediana. Essas observações indicam, no caso de Age, uma distribuição próxima a uma normal, com dados concentrados próximos à média e mediana e uma assimetria à direta no caso de Year of Experience, bem como a possibilidade de outliers à direita.

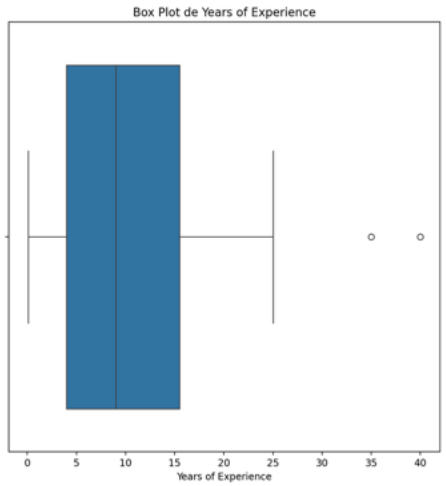
Por fim, o desvio padrão de Age é maior do que o de Years of Experience, indicando um maior espalhamento dos valores com relação à média.

Identificação de Outliers

Age Outliers:



Years of Experience Outliers:



A partir dos box plots pode-se observar, pelo critério de 1,5 IQR, a existência de 2 outliers nos dados de Years of Experience. Os 2 outliers foram retirados para o treinamento do modelo

Dados após limpeza e remoção de outliers

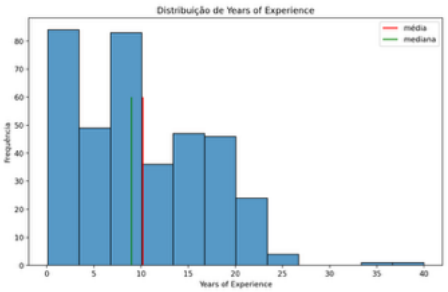
Conforme registrado, foram removidos duas entradas nulas e dois outliers de Years of Experience. A seguir os dados bem como a comparação entre a distribuição de Years of Experience nos 2 momentos.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 373 entries, 0 to 372
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Age                  373 non-null    float64
 1   Gender               373 non-null    object 
 2   Education Level      373 non-null    object 
 3   Job Title            373 non-null    object 
 4   Years of Experience  373 non-null    float64
 5   Salary               373 non-null    float64
```

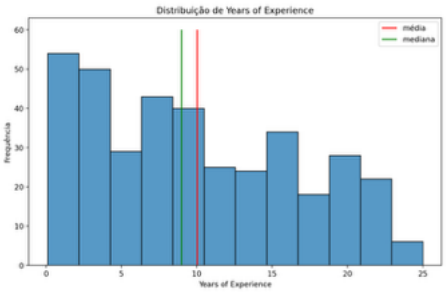
```
dtypes: float64(3), object(3)
memory usage: 17.6+ KB
```

	Age	Gender	Education Level	Job Title	Years of Experience	Salary
0	32	Male	Bachelor's	Software Engineer	5	90,000
1	28	Female	Master's	Data Analyst	3	65,000
2	45	Male	PhD	Senior Manager	15	150,000
3	36	Female	Bachelor's	Sales Associate	7	60,000
4	52	Male	Master's	Director	20	200,000
5	29	Male	Bachelor's	Marketing Analyst	2	55,000
6	42	Female	Master's	Product Manager	12	120,000
7	31	Male	Bachelor's	Sales Manager	4	80,000
8	26	Female	Bachelor's	Marketing Coordinator	1	45,000
9	38	Male	PhD	Senior Scientist	10	110,000

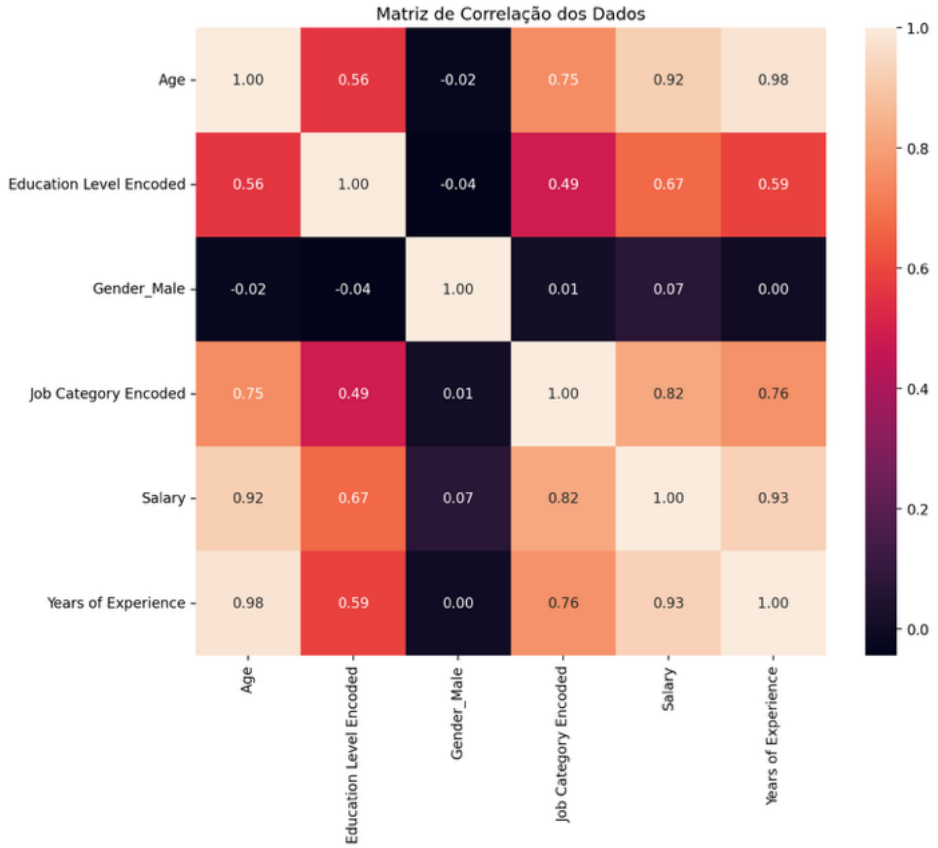
Years of Experience Distribution Inicial:



Years of Experience Distribution Após Ajustes:



Correlação dos Dados com a Variável Alvo (Salary)



Assim observa-se uma forte correlação entre nossa variável alvo (Salary) e Years of Experience, Job Category Encoded, Education Level Encoded e Age. Por outro lado, a variável representativa do gênero (Gender Male) não possui alta correlação com Salary.

Preparação dos Dados

De forma a preparar os dados para a modelagem, foi necessário transformar os dados categóricos em numéricos. Nesse ponto, para o caso de gender, utilizou-se a técnica do One Hot Encoding e excluiu-se uma das resultantes visto que a classificação em homens e mulheres é complementar.

Para o caso de Education Level, utilizou-se o Label Encoding, considerando que existe uma relação de maior valor entre os diferentes níveis de especialização, sendo o mais valioso o PhD.

Por fim, para o caso de Job Title, dividiu-se os dados em 4 categorias utilizando a própria , nomenclatura e também foi executado um Ordinal Encoding considerando a hierarquia entre as posições: (i) , profissionais que possuem junior no nome serão tratados como junior; (ii) profissionais com senior no nome, serão tratados como senior; (iii) profissionais com director, VP e CEO no nome serão tratados como diretores e (iv) outros serão tratados como analitas/plenos, uma faixa intermediária entre o junior e o senior.

Modelagem

Nas modelagens realizadas, não se observou diferença significativa nos resultados devido à aplicação do standard scaler, assim optou-se, considerando o critério de simplicidade, por apresentar o modelo sem a aplicação da transformação. Adicionalmente, em testes realizados observou-se uma melhoria do modelo sem a inclusão da variável age.

Inicialmente optou-se por fazer o teste sem age, pois age e years of experience são duas variáveis altamente correlacionadas. Em modelos de regressão linear variáveis altamente correlacionadas podem trazer ruído para a modelagem, prejudicando o modelo. No caso em análise, observou-se uma melhoria do modelo sem age.

Outro ponto interessante observado foi que em que pese a baixa correlação entre o gênero da pessoa e o salário, o modelo performou melhor com a variável Gender_Male do que sem ela.

Por fim, foram testados diferentes percentuais de split, 30:70, 25:75 e 20:80, sendo o último o que apresentou melhor desempenho. Acreditamos que isso ocorreu devido à baixa quantidade de dados disponíveis.

Nesse sentido, o modelo apresentado a seguir não possui age como feature e foi gerado com um split 20:80.

Resultado da Regressão Linear

Os coeficientes da Regressão Linear são:

value
13,921.9506
6,713.5071
13,079.9714
4,338.2288

O ponto de interseção com o eixo Y é:

`27720.29`

Métricas

	Mean Squared Error	Mean Absolut Error	Root Mean Squared Error	R2 Score	Salário Médio
Valores	184,698,075.72	9,357.83	13,590.37	0.92	100,920.93

Exemplo

Predição do salário de um homem com 15 anos de experiência, em uma posição de senior e com

mestrado:
Salário = 139589.13

Conclusão

Durante a análise realizada, observou-se que os itens que mais influenciam o salário são os anos de experiência, bem como a categoria do trabalho (junior, analista, senior diretor). Por outro lado, o gênero não influenciou o salário. Sendo assim, no caso em análise, seria mais interessante a pessoa se inserir no mercado e começar a trabalhar do que buscar uma maior especialização.

Por fim, foi gerado um modelo que permite predizer o salário de um individuo a partir da apresentação de: (i) nível de educação, (ii) gênero da pessoa, (iii) categoria do trabalho e (iv) anos de experiência.