

Relatório

**icmc
júnior**

Projeto Trainee - Estatística

Sumário

1	Objetivos do projeto	3
2	Desenvolvimento do projeto	3
2.1	Tratamento inicial dos dados	3
2.1.1	Formatar colunas	3
2.1.2	Tratamento de valores nulos	3
2.1.3	Tratamento de variáveis categóricas	3
2.1.4	Tratamento de outliers	4
2.2	Análise Exploratória dos dados	4
2.2.1	Boxplots	4
2.2.2	Histogramas e Densidade	7
2.2.3	Gráficos de frequências	9
2.2.4	Análise de Normalidade	10
2.2.5	Normalização dos Dados	12
2.3	Correlação	12
2.3.1	Matriz de correlação	12
2.3.2	Violin Plots	13
3	Conclusão	15

1 Objetivos do projeto

O objetivo desse projeto é analisar e tratar a base de dados HR Analytics, que reúne informações dos funcionários de uma empresa. Após esse processo, deve ser possível utilizar esses dados em um modelo de machine learning com o objetivo de determinar se um funcionário continuará ou não na empresa. O projeto foi dividido nas seguintes etapas:

- Tratamento inicial dos dados
- Análise Exploratória
- Correlação

2 Desenvolvimento do projeto

O projeto foi feito no Google Notebook utilizando a linguagem Python e suas bibliotecas como Pandas, Matplotlib, Numpy, Sklearn, Seaborn e Scipy. Além disso, parte da análise de dados foi feita utilizando o Power Bi.

2.1 Tratamento inicial dos dados

Essa etapa consiste em tratar os dados de forma a garantir que eles estejam adequados para a análise, isso inclui tratar valores ausentes, formatar colunas e tratar outliers.

2.1.1 Formatar colunas

A primeira etapa foi retirar linhas com valores duplicados, para que não gerem redundâncias nem distorçam métricas como média e medianas. As colunas 'EmployeeCount', 'Over18', 'StandardHours' foram retiradas, pois todas as linhas continham o mesmo valor, e a coluna 'EmployeeNumber' também foi removida.

2.1.2 Tratamento de valores nulos

Em seguida, foi feita uma análise da quantidade de valores ausentes em cada coluna e de métricas básicas como média, mediana e desvio padrão. Esses resultados auxiliaram em uma melhor escolha para a imputação dos dados. Para valores numéricos, foi utilizada a mediana, que não é tão sensível a valores outliers como a média, e, por isso, evita que os dados sofram distorção. Para os valores categóricos, foi utilizada a moda, o valor mais frequente entre todos, para manter a coerência e não introduzir ruído às variáveis.

Por fim, as mesmas métricas, média, mediana e moda, são calculadas a fim de verificar se a imputação causou distorção no conjunto de dados. No projeto, a mediana continuou a mesma, o que era o desejado, e não houve variação significativa na média e no desvio padrão.

2.1.3 Tratamento de variáveis categóricas

O tratamento de variáveis categóricas consiste em transformar essas colunas, normalmente, compostas pelo tipo object para variáveis numéricas, pois alguns modelos podem não saber codificar colunas com valores não numéricos. Para colunas com valores binários como 'Yes' e 'No', cada variável foi transformada em 0 ou 1. Para as demais colunas, foi utilizada a técnica de one-hot encoding que cria uma nova coluna binária para cada categoria. O número de novas colunas criadas não foi tão significativo a ponto de prejudicar a análise do conjunto de dados.

2.1.4 Tratamento de outliers

A análise de outliers consistiu em calcular a porcentagem que eles representam em relação ao todo da coluna e realizar uma imputação ou retirada da coluna. Para identificar os outliers, foi utilizado o método de intervalo de quartis (Q1 e Q3), ou seja, os valores que estão fora do limite inferior e superior dos quartis é um outlier. Dessa maneira, foi possível realizar a porcentagem de outliers em cada coluna.

A coluna 'TrainingTimesLastYear' foi retirada do conjunto, porque os outliers correspondiam a 16% do total dos dados além de não ser uma variável tão importante para o target. Para as demais, foram realizadas duas imputações, uma com a média e outra com a mediana.

Para analisar o resultado das imputações, foram utilizados histogramas e curvas de densidade. Para algumas colunas como 'YearsAtCompany', a distribuição dos dados teve uma mudança significativa para os dois métodos.

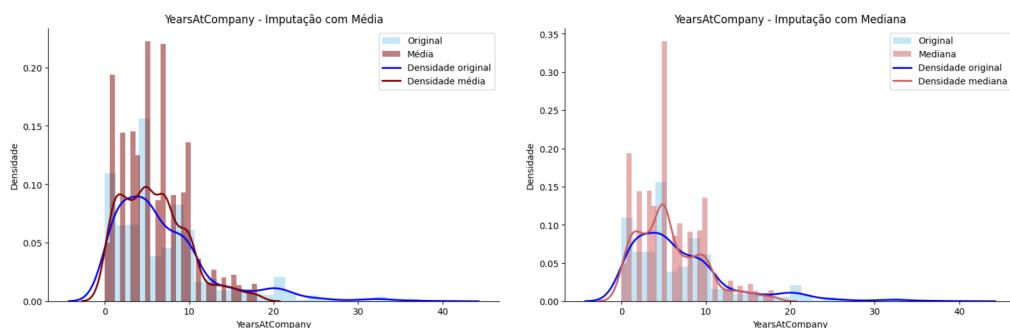


Figura 1: Imputação de outliers

Entretanto, a coluna 'YearsWithCurrManager' apresentou pouca mudança na distribuição dos dados, e o desvio padrão continuou semelhante.

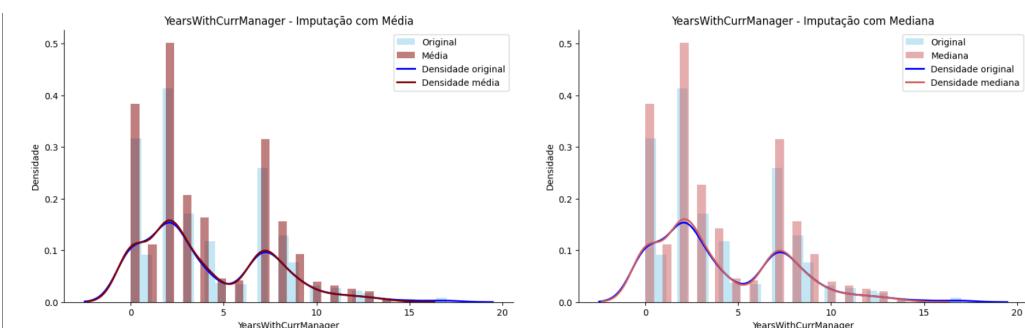


Figura 2: Imputação de outliers

2.2 Análise Exploratória dos dados

Essa próxima etapa consiste em analisar o conjunto de dados para entender mais sobre sua distribuição, normalidade e frequência dos dados.

2.2.1 Boxplots

Os *boxplots* (ou gráficos de caixa) são ferramentas estatísticas essenciais na análise exploratória de dados, especialmente quando lidamos com variáveis numéricas. Eles oferecem uma forma visual clara e concisa de entender a distribuição dos dados, permitindo identificar rapidamente aspectos importantes como a **tendência central**, a **dispersão** e a **presença de outliers** (valores atípicos).

Cada boxplot é construído a partir de cinco números-resumo:

- A **mediana** (linha central da caixa), que representa o valor central da distribuição;
- O **primeiro quartil** (Q1) e o **terceiro quartil** (Q3), que formam os limites inferior e superior da caixa e indicam onde está concentrada metade dos dados;
- Os **limites inferiores e superiores esperados** (whiskers), que geralmente se estendem até 1,5 vezes o intervalo interquartil (IQR) a partir de Q1 e Q3, respectivamente;
- Os **outliers**, ou valores atípicos, que aparecem como pontos isolados fora dos limites esperados e indicam observações que destoam significativamente do padrão da maioria dos dados.

O uso de boxplots é particularmente valioso em contextos onde desejamos:

- Comparar distribuições de diferentes variáveis de forma simultânea;
- Identificar rapidamente assimetrias (distribuições enviesadas), o que pode influenciar decisões sobre o tipo de transformação ou modelo estatístico a ser aplicado;
- Detectar dispersões elevadas e possíveis problemas de escala;
- Evidenciar a existência de valores atípicos que podem afetar a média, o desvio padrão e a performance de modelos preditivos.

No nosso caso específico, a utilização de boxplots foi uma escolha apropriada porque estávamos lidando com múltiplas variáveis numéricas de natureza contínua, como por exemplo *MonthlyIncome*, *TotalWorkingYears* e *YearsAtCompany*. Esses gráficos nos permitiram:

- Ter uma visão global e comparativa da distribuição de cada variável;
- Identificar rapidamente quais variáveis apresentavam maior quantidade de outliers;
- Orientar futuras etapas do processo de pré-processamento de dados, como normalização, transformação ou tratamento de valores extremos.

Portanto, o uso de boxplots não apenas enriqueceu nossa análise exploratória, como também forneceu subsídios importantes para decisões posteriores no pipeline de Ciência de Dados, contribuindo para maior robustez e confiabilidade dos resultados. Conforme a figura 1 é possível ver como os dados foram distribuídos nos gráficos.

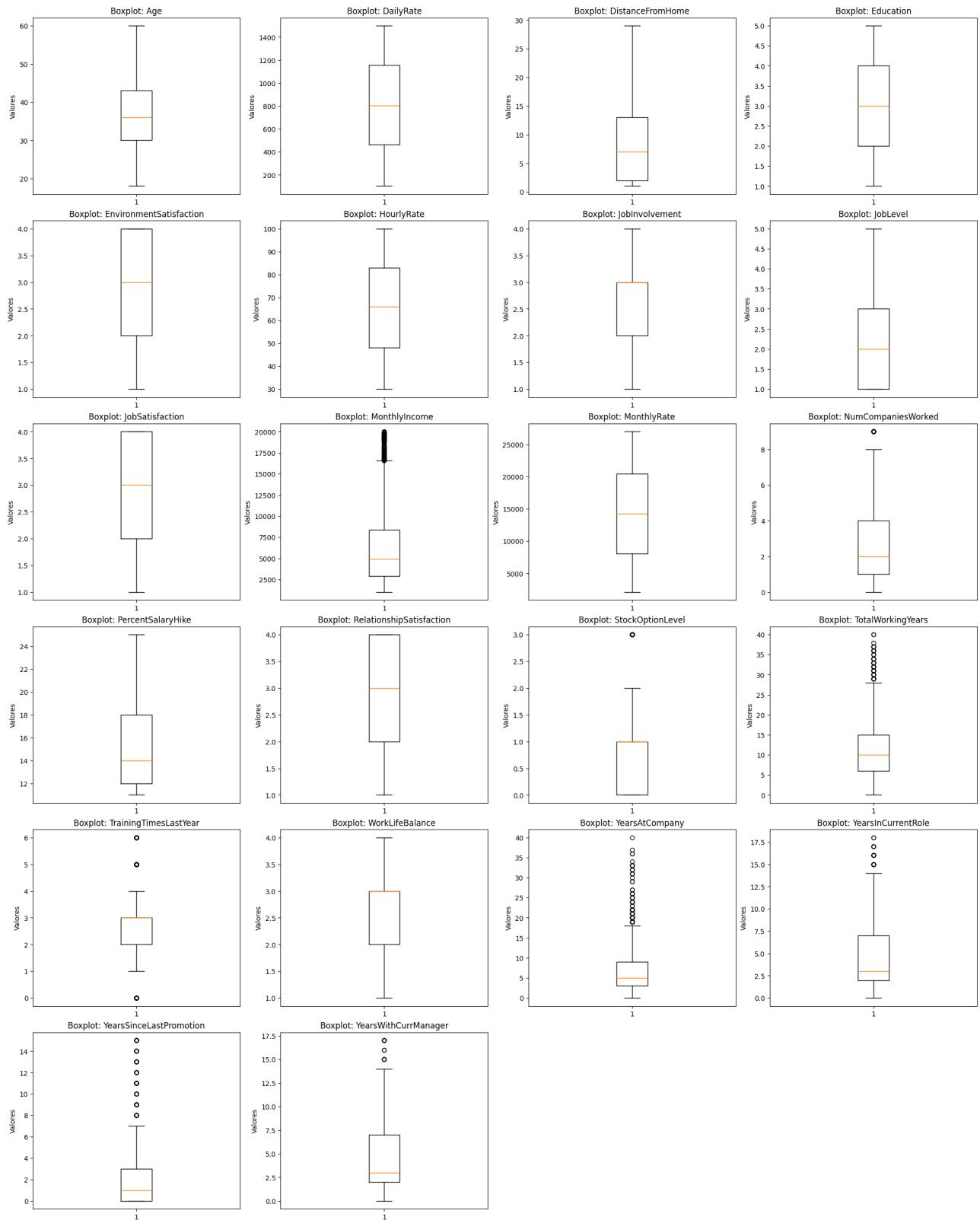


Figura 3: BoxPlot

Durante a análise, observou-se que as variáveis:

- `MonthlyIncome` (salário mensal),
- `TotalWorkingYears` (tempo total de experiência),

- YearsAtCompany (tempo de empresa),
- YearsSincePromotion (tempo desde a última promoção)

apresentaram **grande concentração de outliers**, o que indica que há funcionários com valores significativamente acima ou abaixo do padrão observado para a maioria.

Esse achado é relevante, pois:

- Pode indicar a existência de **grupos específicos** (ex: veteranos ou recém-contratados);
- Reforça a necessidade de **tratamento cuidadoso desses valores** antes da modelagem, seja por meio de normalização robusta ou imputação;
- Alerta para possíveis **distorções estatísticas**, caso essas variáveis não sejam tratadas adequadamente.

Assim, os *boxplots* mostraram-se uma ferramenta essencial para compreender a estrutura dos dados e orientar decisões nas etapas seguintes da análise.

2.2.2 Histogramas e Densidade

Os histogramas são representações gráficas que mostram a distribuição de uma variável numérica, dividindo os dados em intervalos (ou “bins”) e exibindo a frequência de ocorrência de valores em cada intervalo. Esse tipo de gráfico é extremamente útil na análise exploratória de dados, pois permite uma compreensão visual clara sobre o formato da distribuição: se ela é simétrica, assimétrica, uniforme, multimodal ou se apresenta caudas longas.

Além disso, os histogramas ajudam a identificar:

- A presença de **distribuições normais ou não normais**, o que é fundamental para a escolha de testes estatísticos e modelos apropriados;
- **Concentrações de dados** em determinados intervalos;
- **Lacunas (gaps)** ou **valores extremos** que, embora não apareçam como outliers isolados, podem distorcer a distribuição geral;
- **Tendências gerais** como inclinação para a esquerda (assimetria negativa) ou para a direita (assimetria positiva).

No caso específico do projeto, a utilização de histogramas foi uma ferramenta importante para avaliar visualmente se as variáveis numéricas seguiam uma distribuição aproximadamente normal, o que influencia diretamente em decisões posteriores como a escolha de métodos de normalização, padronização e a aplicabilidade de modelos estatísticos que assumem normalidade. Como é possível observar nos gráficos da figura 4:

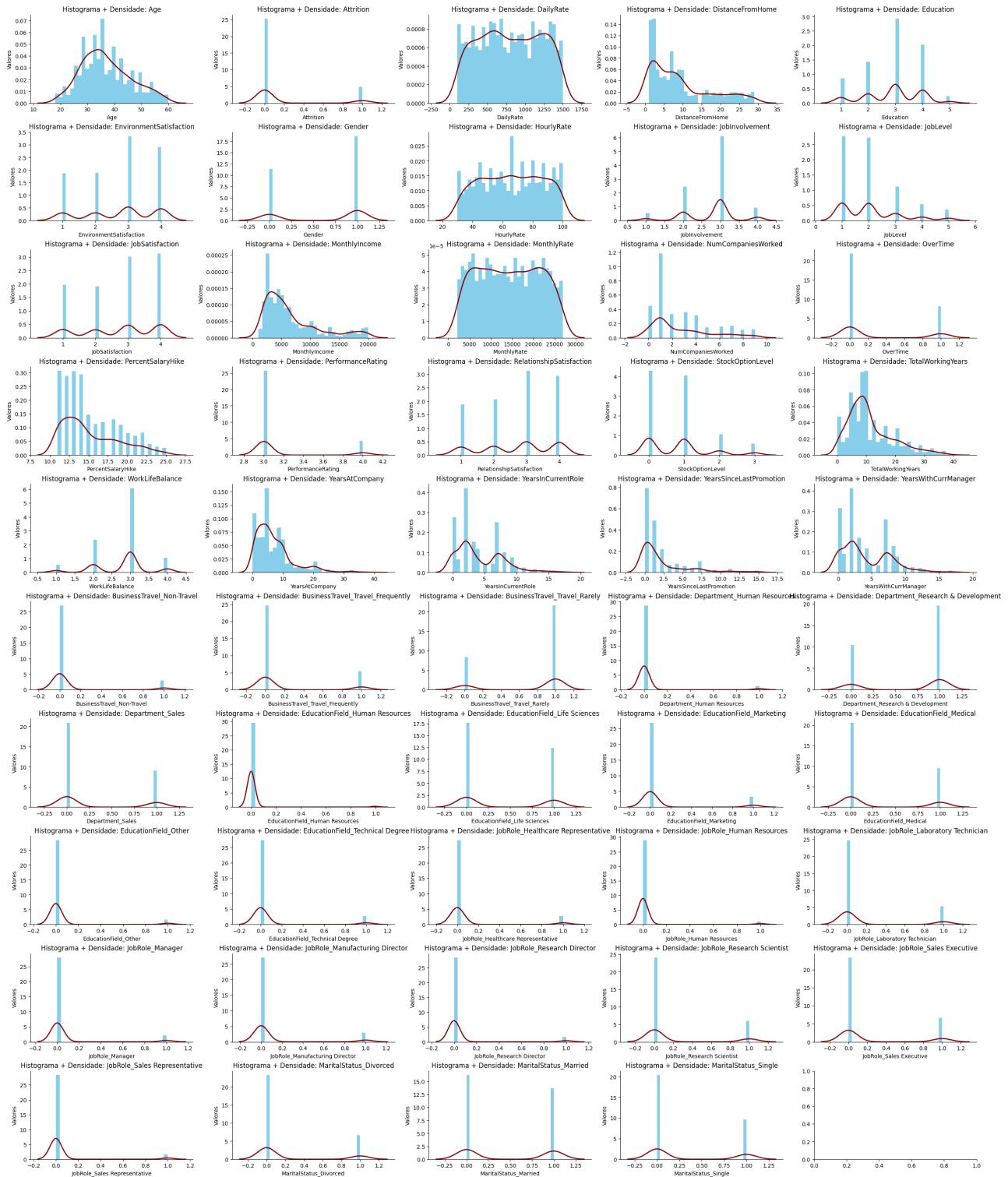


Figura 4: Histogramas e densidaade

A análise dos histogramas revelou padrões relevantes sobre a distribuição das variáveis numéricas da base de dados. De forma geral, observou-se que muitas variáveis não seguem uma distribuição normal, sendo assimétricas ou concentradas em intervalos específicos. Essa constatação foi essencial para orientar decisões de pré-processamento, especialmente na escolha dos métodos de normalização.

As principais observações feitas a partir dos histogramas incluem:

- A variável **YearsAtCompany** apresentou forte concentração de funcionários com pouco tempo de empresa, com frequência decrescendo à medida que o tempo aumenta. Isso indica uma possível alta rotatividade ou entrada recente de colaboradores;
- De maneira semelhante, **YearsInCurrentRole** mostrou que a maioria dos funcionários está há pouco tempo na função atual, sugerindo movimentações internas frequentes ou crescimento recente da equipe;
- A variável **MonthlyIncome** exibiu distribuição assimétrica com cauda à direita, indicando que a maior parte dos funcionários recebe salários mais baixos, com poucos indivíduos recebendo valores mais altos — comportamento típico de variáveis salariais;
- A distribuição de **TotalWorkingYears** também apresentou assimetria positiva, indicando predominância de colaboradores com menos anos de experiência profissional no geral;
- A variável **YearsWithCurrManager**, usada na análise de imputação, demonstrou baixa variação entre os métodos (média e mediana), com distribuições visivelmente semelhantes antes e depois do tratamento, reforçando a consistência dos dados;
- Já variáveis como **PercentSalaryHike** e **PerformanceRating** apresentaram distribuições com pouca variabilidade — ou seja, muitos valores concentrados em uma ou poucas faixas. Isso sugere que há pouca diferenciação entre os funcionários nesses critérios, o que pode reduzir o poder preditivo dessas variáveis;
- Os histogramas também serviram como ferramenta de apoio na avaliação do tratamento de outliers. Após a imputação, foi possível visualizar que, em variáveis como **YearsAtCompany**, a distribuição foi significativamente alterada, indicando sensibilidade à técnica aplicada, enquanto em outras, como **YearsWithCurrManager**, a mudança foi mínima.

Essas análises reforçaram a decisão de utilizar métodos de normalização distintos, como o *MinMaxScaler* para variáveis com distribuição mais regular e sem outliers, e o *RobustScaler* para aquelas com assimetria acentuada e presença de valores extremos.

Tabela 1: Resumo das observações a partir dos histogramas

Variável	Padrão observado	Interpretação / Impacto
YearsAtCompany	Concentração nos valores baixos	Alta rotatividade ou muitos funcionários recentes
YearsInCurrentRole	Concentração próxima de zero	Muitas pessoas em funções novas ou realocadas recentemente
MonthlyIncome	Assimetria positiva (cauda à direita)	Poucos com salário alto, maioria com salários menores — típico de distribuição salarial
TotalWorkingYears	Assimetria positiva	Predominância de profissionais com pouca experiência
PercentSalaryHike	Distribuição com baixa variabilidade	Diferença salarial entre funcionários é pequena
PerformanceRating	Distribuição concentrada em poucos valores	Avaliação de desempenho pouco variada, todos avaliados de forma semelhante

2.2.3 Gráficos de frequências

Os gráficos de frequência são representações visuais utilizadas para exibir a quantidade de ocorrências (frequência absoluta ou relativa) de cada categoria presente em variáveis qualitativas. Eles são especialmente úteis para variáveis categóricas, pois permitem uma análise clara e objetiva da distribuição dos dados entre as diferentes classes.

Esses gráficos tornam possível:

- Visualizar de forma imediata **quais categorias são mais ou menos frequentes**;
- **Comparar a proporção entre categorias**, mesmo quando há muitas classes;
- Identificar **desequilíbrios na distribuição**, como concentração excessiva em uma única categoria;
- Ajudar na **tomada de decisão sobre reagrupamento ou transformação de categorias**, especialmente quando se trabalha com algoritmos sensíveis ao desequilíbrio.

No nosso caso, os gráficos de frequência foram aplicados exclusivamente às variáveis categóricas, como por exemplo gênero, cargo e estado civil. Essa visualização nos possibilitou identificar rapidamente a quantidade de registros associados a cada categoria, o que é fundamental para entender a estrutura do conjunto de dados e planejar etapas posteriores de pré-processamento, como codificação (one-hot encoding, label encoding) e tratamento de valores raros.

- Em **JobRole**, algumas funções como *Sales Executive* e *Research Scientist* estavam significativamente mais representadas do que outras, como *Healthcare Representative*;
- Em **MaritalStatus**, a categoria *Married* era a mais frequente, seguida de *Single*, com uma quantidade consideravelmente menor de registros para *Divorced*;
- Já na variável **Gender**, observamos uma distribuição relativamente equilibrada entre homens e mulheres.

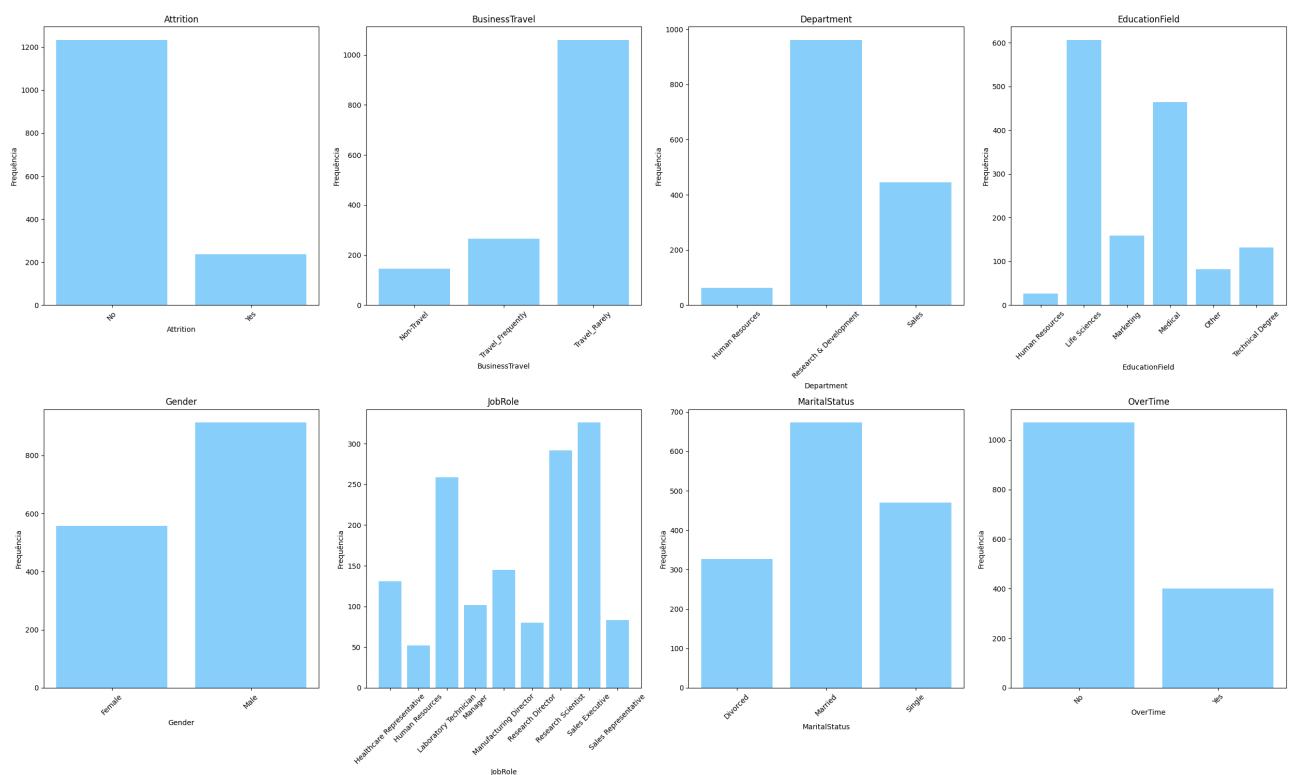


Figura 5: Gráficos de Frequência

2.2.4 Análise de Normalidade

Essa etapa consiste em analisar se os dados possuem ou não uma distribuição normal. Por meio dos histogramas, é possível verificar visualmente que em geral os dados não estão distribuídos normalmente. Mesmo assim, foram feitos mais gráficos do tipo QQ e foram realizados testes de Shapiro-Wilk, Anderson-Darling e D'Agostino-Pearson, que foram escolhidos considerando o volume dos dados e a distribuição dos dados.

Os gráficos mostram que os dados não possuem uma distribuição normal:

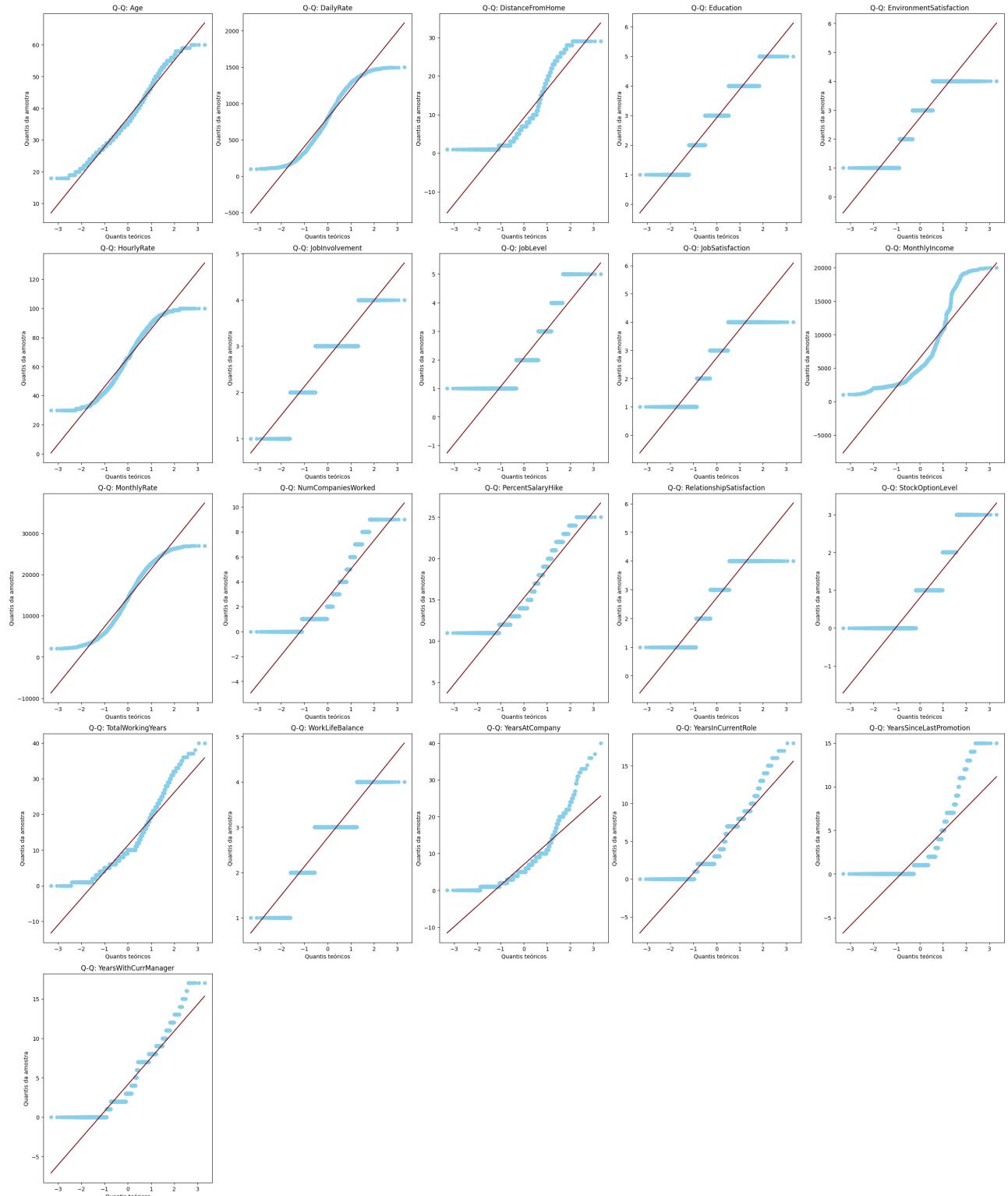


Figura 6: QQ

Além disso, todas as colunas não possuem distribuição normal para todos os testes realizados (Shapiro-Wilk, Anderson-Darling e D'Agostino-Pearson).

2.2.5 Normalização dos Dados

A normalização dos dados é importante para colocar os dados de escalas diferentes na mesma escala de forma que não haja uma dominação no modelo de algumas variáveis. Além disso, quando os dados estão normalizados, a interpretação e desempenho do modelo ficam melhores. Para a escolha de qual método de normalização seria usado, foram levados em consideração a quantidade de outliers e o tipo de distribuição dos dados. Para as colunas que não possuem outliers, foi usado o método MinMaxScaler, já que ele é sensível a amostras com muitos outliers. Já para as outras colunas, foi usado o método RobustScaler que é mais robusto em questão da quantidade de outliers e também para dados que não possuem distribuição normal, como é o caso.

2.3 Correlação

A matriz de correlação foi utilizada nesta etapa porque a correlação é uma ferramenta estatística que nos permite medir o grau de associação entre duas variáveis numéricas. Em outras palavras, ela nos mostra se existe alguma relação entre duas variáveis e qual a direção e intensidade dessa relação.

Por exemplo, ao calcular a correlação entre o número de anos que um funcionário está na empresa e sua chance de sair (variável Attrition), conseguimos observar se há uma tendência de funcionários com menos tempo de casa estarem mais propensos a sair. Se existir essa tendência, a correlação será negativa (valores próximos de -1), indicando que conforme uma variável aumenta, a outra tende a diminuir.

A principal razão para aplicar a matriz de correlação neste projeto é identificar quais variáveis possuem algum tipo de relação com a variável Attrition, nosso foco de interesse. Assim, conseguimos destacar os fatores que mais influenciam a permanência ou saída de um funcionário da empresa.

Além disso, a matriz de correlação nos ajuda a perceber outras relações entre variáveis do conjunto de dados. Por exemplo, se duas variáveis como MonthlyIncome e JobLevel estiverem muito correlacionadas entre si, pode não ser necessário manter as duas em análises futuras, já que elas carregam informações redundantes. Isso é importante porque variáveis altamente correlacionadas entre si (colinearidade) podem prejudicar o desempenho e a interpretação de modelos estatísticos ou de machine learning.

Portanto, essa etapa é essencial para:

- Identificar variáveis que estão mais associadas à saída dos funcionários (positiva ou negativamente);
- Guiar a seleção de variáveis relevantes para etapas seguintes de análise e modelagem;
- Detectar relações redundantes entre variáveis, que podem ser descartadas para evitar interferência nos resultados.

Em resumo, utilizamos a matriz de correlação porque ela é uma ferramenta poderosa para encontrar padrões escondidos nos dados, nos permitindo tomar decisões mais informadas e eficazes nas etapas posteriores do projeto.

2.3.1 Matriz de correlação

Para aplicar a matriz de correlação na prática, utilizamos uma função em Python que recebe como entrada o *Dataframe* com os dados numéricos e o método desejado para o cálculo da correlação (como o método de Spearman, que foi o utilizado neste caso). Internamente, a função utiliza o comando `df.corr(method='spearman')` para calcular os coeficientes de correlação entre todas as variáveis numéricas do conjunto de dados. Esse método foi escolhido, pois é mais adequado para a distribuição não normal dos dados, como é o caso.

Esses coeficientes foram então representados visualmente por meio de um *heatmap* (mapa de calor), gerado com a biblioteca Seaborn. O gráfico utiliza uma escala de cores (`coolwarm`) para indicar o grau de correlação entre as variáveis: cores próximas do azul indicam correlação negativa, próximas do vermelho indicam correlação positiva, e tons neutros indicam correlação fraca ou nula. Os valores numéricos também foram exibidos dentro de cada célula, facilitando a interpretação visual da matriz.

Essa visualização permitiu identificar rapidamente as variáveis com maior associação com a variável Attrition, direcionando as análises seguintes e contribuindo para uma compreensão mais clara das relações existentes nos dados.

Essa função recebe um DataFrame e o método de correlação (Pearson, Spearman, etc.) como entrada, e gera um heatmap que permite visualizar graficamente as correlações entre todas as variáveis numéricas.

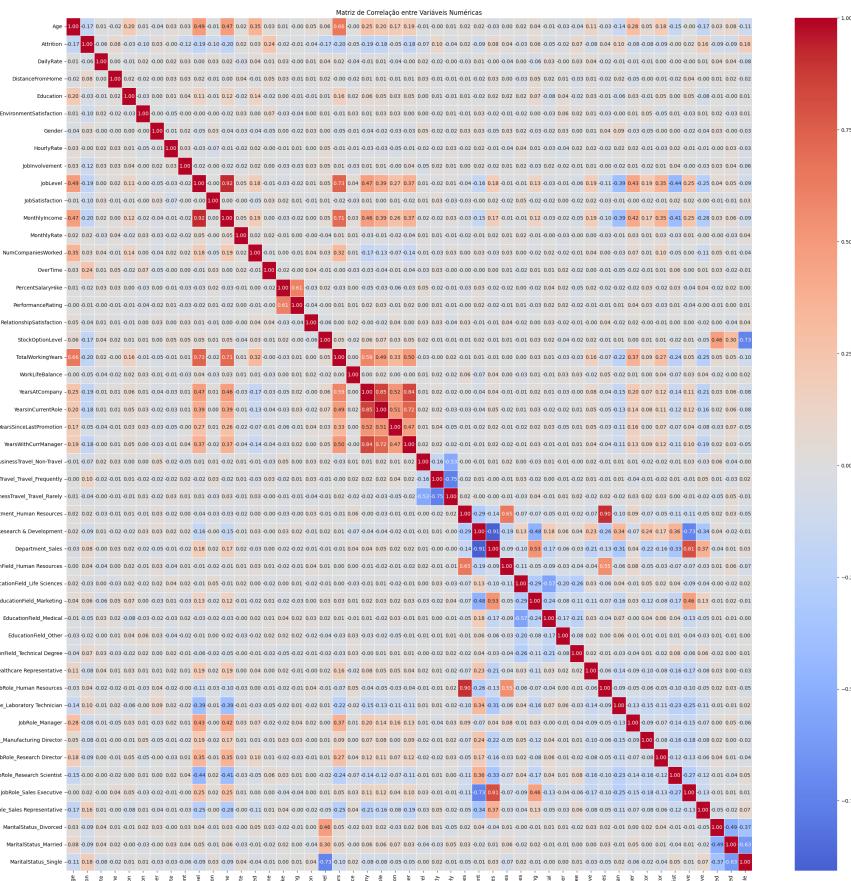


Figura 7: Matriz de correlaco

Observe que as variáveis que mais se correlacionam com a variável target ('Attrition') são 'OverTime', 'JobRole SalesRepresentative' e 'MaritalStatus Single'.

2.3.2 Violin Plots

Para complementar a análise visual, foram gerados violin plots para cada variável numérica em relação à variável Attrition. Esse tipo de gráfico combina elementos de um boxplot tradicional — como mediana, quartis e outliers — com a curva de densidade estimada, permitindo uma visualização mais rica e intuitiva da distribuição dos dados.

Diferentemente dos boxplots, que mostram apenas estatísticas-resumo, os violin plots também revelam o formato da distribuição (simétrica, assimétrica, unimodal ou multimodal), o que pode indicar agrupamentos ou padrões relevantes.

Além disso, ao observar a largura da densidade em diferentes faixas de valores, é possível perceber com mais clareza como as variáveis numéricas se distribuem entre os grupos com e sem evasão (Attrition = Yes ou No). Isso facilita a identificação de variáveis potencialmente correlacionadas com a saída dos funcionários, servindo como guia para etapas futuras de modelagem e seleção de variáveis.

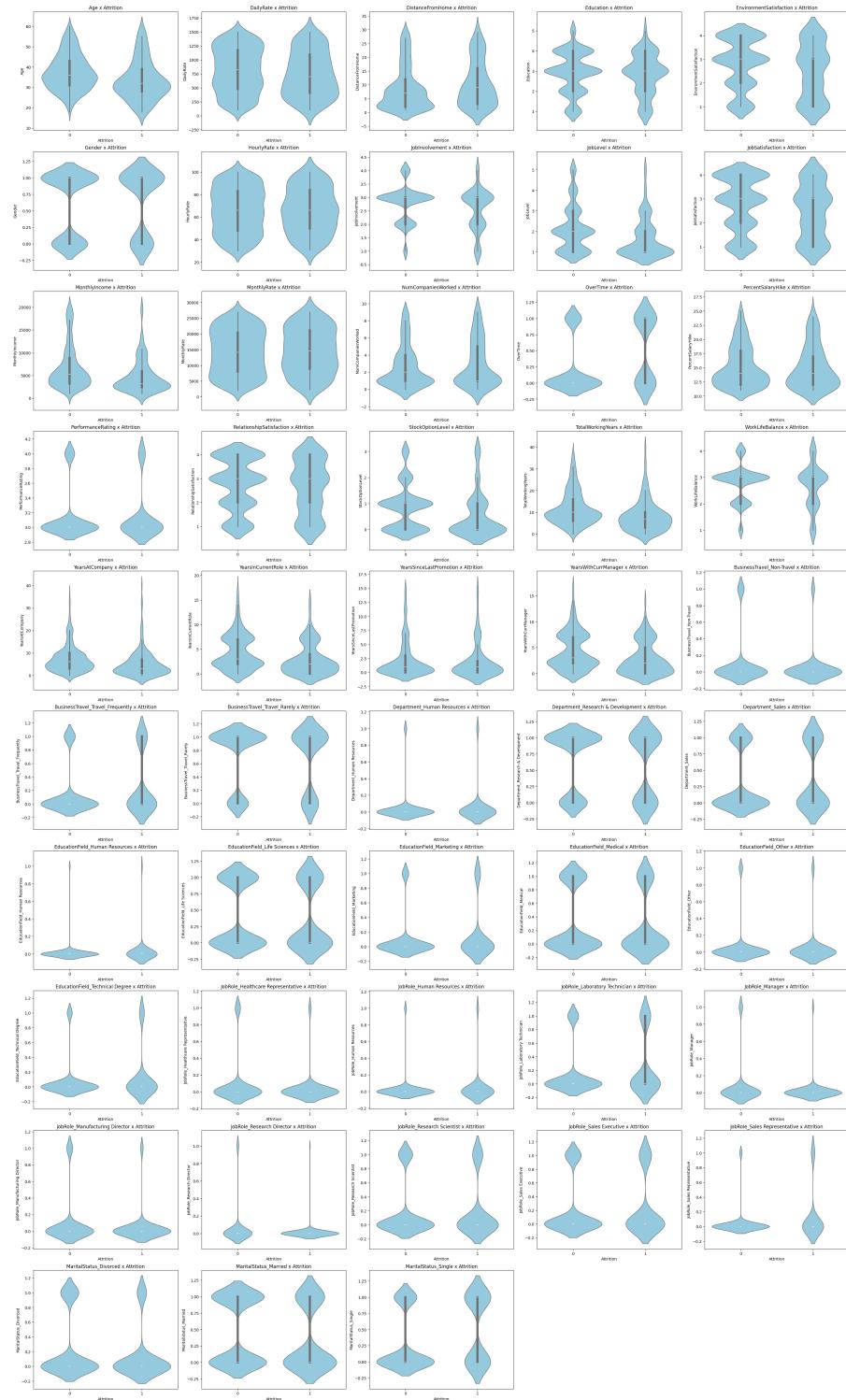


Figura 8: Violin Plots

A análise dos violin plots proporcionou conclusões sobre a distribuição das variáveis numéricas em relação à evasão de funcionários (Attrition). Por meio dessa visualização, foi possível identificar com maior clareza quais grupos de colaboradores apresentam maior tendência de saída da empresa. As principais conclusões obtidas foram:

- **Horas Extras (OverTime):** Funcionários que realizam horas extras apresentaram maior densidade no grupo com evasão (Attrition = Yes). Esse padrão sugere que a sobrecarga de trabalho pode estar associada a um aumento do estresse ou insatisfação, levando à saída da empresa.

- **Salário Mensal (MonthlyIncome):** Observou-se que os funcionários com salários mais baixos concentram-se majoritariamente entre os que deixaram a empresa. Isso indica uma possível relação entre menor remuneração e maior rotatividade, refletindo insatisfação ou busca por melhores oportunidades.
- **Idade (Age):** A densidade de funcionários mais jovens é maior no grupo que saiu, o que pode indicar menor estabilidade no início da carreira ou maior disposição para mudanças e crescimento profissional em outras empresas.
- **Tempo de Empresa e Experiência (YearsAtCompany, TotalWorkingYears):** Funcionários com pouco tempo de casa e menor experiência geral mostraram maior propensão à saída, sugerindo que retenção pode estar associada ao amadurecimento profissional e integração organizacional ao longo do tempo.

3 Conclusão

Neste projeto, realizamos uma análise exploratória inicial dos dados relacionados à evasão de funcionários (Attrition), aplicando técnicas fundamentais de pré-processamento, como a formatação de colunas, tratamento de valores ausentes e codificação de variáveis categóricas, que possibilitaram a estruturação adequada da base para análises posteriores.

A etapa central consistiu na investigação estatística das variáveis numéricas por meio da matriz de correlação e de representações gráficas como os violin plots. Dentre os métodos aplicados, a análise de correlação foi essencial para identificar quais fatores estão mais associados à decisão dos funcionários de deixarem a empresa, permitindo um entendimento quantitativo das relações entre as variáveis.

A Tabela abaixo resume os principais resultados obtidos:

Variável	Correlação	Interpretação
OverTime	+0.24	Funcionários que fazem mais horas extras tendem a sair com mais frequência.
JobSatisfaction	-0.21	Menor satisfação no trabalho → maior chance de saída.
MonthlyIncome	-0.17	Salários mais baixos estão associados com maior rotatividade.
YearsAtCompany	-0.13	Funcionários com pouco tempo de casa tendem a sair mais.
TotalWorkingYears	-0.12	Menos experiência geral → mais propensos a sair.
YearsInCurrentRole	-0.11	Recém-alocados no cargo atual saem com mais facilidade.
DistanceFromHome	+0.11	Morar longe influencia negativamente na permanência.
Age	-0.10	Funcionários mais jovens tendem a sair com mais frequência.

Tabela 2: Correlações com a variável Attrition

Destacam-se como fatores com correlação positiva com a evasão o trabalho em horário extra (OverTime), a distância até a empresa e a pouca experiência no cargo atual. Por outro lado, variáveis como satisfação no trabalho, salário mensal, tempo na empresa e idade apresentaram correlação negativa, ou seja, quanto menores esses valores, maior a chance de evasão.

Embora os coeficientes de correlação encontrados sejam, em sua maioria, de baixa magnitude (inferiores a $|0.25|$), eles oferecem indícios relevantes para etapas futuras de modelagem e tomada de decisão. Esses resultados indicam que a evasão é influenciada por múltiplos fatores que, apesar de individualmente fracos, podem atuar em conjunto para compor um perfil de risco.

Portanto, a análise realizada fornece uma base sólida para o aprofundamento do estudo, sendo especialmente útil na seleção de variáveis preditivas para modelos supervisionados e no direcionamento de estratégias de retenção mais alinhadas com os dados observados.