PUC-RJ

Pós Graduação em Ciência de Dados e Analytics

Vanessa Araújo da Silva

MVP PIPELINE DE DADOS

Recursos Humanos

Análise de Retenção e Eficiência Organizacional

Belo Horizonte

2025

1. Escopo do Trabalho

1.1. Definição do Problema

O presente trabalho tem como objetivo responder a questionamentos de negócio relacionados à **gestão de pessoas**, utilizando dados de Recursos Humanos para apoiar a tomada de decisão estratégica em uma organização de médio porte. Em ambientes organizacionais, especialmente aqueles com operações industriais ou operacionais, fatores como **rotatividade de colaboradores, absenteísmo, desempenho, engajamento e diversidade** impactam diretamente a eficiência, os custos operacionais e a sustentabilidade do negócio.

A empresa analisada enfrenta desafios relacionados à **retenção de talentos**, ao **alto índice de ausências em determinados departamentos** e à necessidade de compreender melhor os fatores que influenciam o desempenho e a satisfação dos colaboradores. A ausência de uma visão integrada e estruturada dos dados de RH dificulta a identificação de padrões e gargalos organizacionais, limitando a atuação preventiva da gestão.

Dessa forma, este trabalho propõe a utilização de uma **base de dados pública e anonimizada de Recursos Humanos**, que será detalhada nas seções seguintes, com o objetivo de construir um pipeline de dados e responder às seguintes questões de negócio:

a) Quais fatores estão mais associados à rotatividade de colaboradores?
Busca-se identificar padrões relacionados a desligamentos voluntários ou involuntários, analisando variáveis como departamento, tempo de empresa, desempenho, engajamento e fonte de recrutamento. A compreensão desses fatores pode auxiliar a organização a desenvolver estratégias mais eficazes de retenção de talentos.

b) Quais departamentos apresentam maiores índices de absenteísmo e quais os possíveis impactos organizacionais?
A análise do volume de ausências por departamento permite identificar áreas críticas, como setores operacionais ou de produção, onde o absenteísmo pode afetar diretamente a produtividade e a qualidade dos processos. Essa informação é essencial para a proposição de ações corretivas e preventivas.

c) Existe relação entre desempenho, engajamento e tempo de empresa?
Ao analisar avaliações de desempenho, pesquisas de engajamento e tempo de permanência na organização, é possível compreender se colaboradores mais engajados apresentam melhor desempenho e maior retenção, fornecendo subsídios para políticas de desenvolvimento e clima organizacional.

d) Como a diversidade está distribuída entre os departamentos e qual seu impacto na retenção?
A análise da distribuição racial e de diversidade permite avaliar se existem desequilíbrios entre áreas da empresa e se tais fatores influenciam indicadores como desligamento e engajamento, contribuindo para a formulação de políticas de inclusão.

Essas análises permitem que a organização utilize dados de RH de forma estratégica, reduzindo custos associados à rotatividade, melhorando a eficiência operacional e promovendo um ambiente de trabalho mais sustentável.

1.2. Criação de uma Pipeline de Dados

Para dar suporte às análises propostas, será construída uma **pipeline de dados em ambiente de computação em nuvem**, utilizando a plataforma **Databricks Community Edition**, que é gratuita e amplamente utilizada em projetos de Engenharia de Dados e Analytics. Apesar de possuir limitações de recursos computacionais por se tratar de uma versão free, a plataforma é adequada para o desenvolvimento do MVP proposto.

O Databricks é uma plataforma de dados integrada que oferece suporte à ingestão, processamento, transformação e análise de dados, permitindo a implementação de pipelines de **ETL (Extração, Transformação e Carga)** de forma escalável e organizada. No contexto deste trabalho, a plataforma será utilizada para:

- Armazenar os dados brutos de Recursos Humanos;
- Realizar transformações e tratamentos necessários para padronização e qualidade dos dados;
- Estruturar os dados em camadas analíticas adequadas para análise;
- Apoiar a execução de consultas analíticas por meio de SQL e/ou Python.

A utilização de uma pipeline de dados estruturada possibilita maior rastreabilidade, reprodutibilidade das análises e melhor governança dos dados, aspectos fundamentais em projetos de Engenharia de Dados.

---

2. Fonte de Dados

2.1. Descrição da Fonte de Dados

A fonte de dados utilizada neste trabalho foi obtida a partir da plataforma **Kaggle**, um repositório amplamente reconhecido por profissionais das áreas de Ciência de Dados, Engenharia de Dados e Analytics. O Kaggle disponibiliza conjuntos de dados públicos e de alta qualidade, utilizados tanto para fins acadêmicos quanto profissionais.

O conjunto de dados selecionado denomina-se **"Base de Dados Recursos Humanos (RH)"**, disponibilizado em versão tratada e traduzida, com origem no dataset original *Human Resources Data Set*, criado pela Dra. Carla Patalano e pelo Dr. Rich Huebner. A versão utilizada passou por processos de limpeza e padronização para fins exploratórios e analíticos.

De acordo com a documentação disponibilizada pelos autores, o conjunto de dados contém informações anonimizadas sobre colaboradores, incluindo dados demográficos, departamento de atuação, tempo de empresa, desempenho, engajamento, ausências, diversidade e informações relacionadas a desligamentos. Esses dados permitem análises voltadas à compreensão de fatores que impactam a retenção de talentos e a eficiência organizacional.

O conjunto de dados está licenciado sob a licença **Creative Commons Atribuição-NãoComercial-SemDerivações 4.0 Internacional**, sendo permitido seu uso para fins acadêmicos, conforme proposto neste trabalho.

A escolha desta base de dados se justifica por sua aderência ao problema definido, pela riqueza de atributos relacionados à gestão de pessoas e pela possibilidade de simular um cenário realista de análise de dados de RH, alinhado aos conceitos de Engenharia de Dados e pipelines em nuvem.

2.2. Features da Fonte de Dados

Nesta seção são detalhados os campos, ou *features*, do conjunto de dados de **Recursos Humanos (RH)** obtido a partir da plataforma Kaggle e utilizado para o desenvolvimento deste trabalho.

As principais *features* utilizadas neste trabalho são descritas a seguir:

1. **ID do Funcionário**: Identificador único e anonimizado de cada colaborador no conjunto de dados, utilizado para controle e relacionamento entre tabelas.

2. **Departamento**: Área organizacional na qual o colaborador atua (por exemplo, Produção, Vendas, RH, TI, entre outros).

3. **Cargo**: Função ou posição ocupada pelo colaborador dentro da organização.

4. **Idade**: Idade do colaborador no momento do registro dos dados.

5. **Gênero**: Informação referente ao gênero do colaborador, utilizada para análises demográficas.

6. **Raça/Etnia**: Classificação racial do colaborador, utilizada para análises de diversidade e inclusão.

7. **Estado Civil**: Situação civil do colaborador, podendo ser utilizada como variável complementar em análises demográficas.

8. **Data de Admissão**: Data de entrada do colaborador na organização.

9. **Data de Desligamento**: Data de saída do colaborador da organização, quando aplicável, permitindo a análise de rotatividade.

10. **Status do Funcionário**: Indica se o colaborador encontra-se ativo ou desligado da empresa.

11. **Tempo de Empresa**: Tempo total de permanência do colaborador na organização, geralmente expresso em anos, calculado a partir das datas de admissão e desligamento.

12. **Fonte de Recrutamento**: Canal pelo qual o colaborador foi contratado (por exemplo, indicação, site de vagas, agência externa), permitindo avaliar a efetividade das estratégias de recrutamento.

13. **Avaliação de Desempenho**: Indicador qualitativo ou quantitativo que representa o desempenho do colaborador em avaliações periódicas.

14. **Pontuação de Engajamento**: Resultado de pesquisas internas de engajamento ou satisfação dos colaboradores.

15. **Pesquisa GPTW (Great Place to Work)**: Indicador relacionado à percepção do colaborador sobre o ambiente de trabalho.

16. **Dias de Ausência**: Quantidade de dias de ausência registrados em um determinado período.

17. **Atrasos**: Número de registros de atrasos do colaborador, utilizado como indicador complementar de comportamento organizacional.

18. **ID do Gestor**: Identificador do gestor imediato do colaborador, utilizado para análises hierárquicas e de liderança.

19. **Nome do Gestor**: Informação textual referente ao gestor direto, utilizada para facilitar a interpretação dos dados.

20. **Salário Anual**: Valor do salário anual do colaborador. Neste trabalho, essa informação é utilizada apenas de forma agregada, respeitando o caráter confidencial dos dados.

Essas *features* permitem a construção de métricas e indicadores estratégicos relacionados à **rotatividade, absenteísmo, desempenho, engajamento, diversidade e custos de pessoal**, sendo fundamentais para o desenvolvimento das análises propostas neste MVP. Além disso, a variedade de atributos possibilita a estruturação dos dados em modelos analíticos adequados, como esquemas estrela ou camadas de Data Lake, conforme será apresentado nas seções seguintes.

3. Modelo e Catálogo de Dados

3.1. Modelagem de Dados

A partir da base de dados de Recursos Humanos obtida no Kaggle, foi construído um modelo de dados seguindo o **Esquema Estrela**, amplamente utilizado em projetos de Data Warehousing e Analytics.

O Esquema Estrela é caracterizado por uma **tabela de fatos central**, que armazena métricas quantitativas do negócio, e por **tabelas de dimensões**, que descrevem os contextos nos quais essas métricas são analisadas.

Neste projeto, a tabela de fatos central é a **fact_employee**, e as tabelas de dimensões são **dim_employee**, **dim_department**, **dim_position** e **dim_manager**, conforme descrito a seguir:

- **Tabela de Fatos “fact_employee”**:
  Contém métricas relacionadas aos funcionários, como salário, ausências, atrasos, satisfação, desempenho e datas de contratação e desligamento, além das chaves estrangeiras para as tabelas de dimensão.

- **Tabela de Dimensão “dim_employee”**:
  Armazena informações demográficas e pessoais dos funcionários, como gênero, raça, estado civil e cidadania.

- **Tabela de Dimensão “dim_department”**:
  Armazena informações relacionadas aos departamentos da organização.

- **Tabela de Dimensão “dim_position”**:
  Armazena informações sobre os cargos ocupados pelos funcionários.

- **Tabela de Dimensão “dim_manager”**:
  Armazena informações sobre os gestores responsáveis pelos funcionários.

3.2. Catálogo de Dados

A seguir, apresenta-se o **catálogo de dados**, contendo a descrição detalhada das tabelas do esquema Silver, incluindo tipos de dados, chaves primárias, chaves estrangeiras e domínios esperados.

A. Tabela de Fatos “silver.fact_employee”

1. **employee_id (PK, STRING)**
   Identificador único do funcionário.

2. **department_id (FK, STRING)**
   Chave estrangeira referenciando a tabela dim_department.

3. **position_id (FK, STRING)**
   Chave estrangeira referenciando a tabela dim_position.

4. **manager_id (FK, STRING)**
   Chave estrangeira referenciando a tabela dim_manager.

5. **hire_date (DATE)**
   Data de contratação do funcionário.

6. **termination_date (DATE)**
   Data de desligamento do funcionário, quando aplicável.

7. **salary (DOUBLE)**
   Salário anual do funcionário em dólares americanos.

8. **absences (INTEGER)**
   Número total de ausências registradas.

9. **days_late_last_30 (INTEGER)**
   Quantidade de atrasos nos últimos 30 dias.

10. **performance_score (STRING)**
    Avaliação de desempenho do funcionário (ex.: Excelente, Atende, PIP).

11. **employee_satisfaction (INTEGER)**
    Nível de satisfação do funcionário, variando tipicamente de 1 a 5.

12. **special_projects_count (INTEGER)**
    Quantidade de projetos especiais em que o funcionário participou.

B. Tabela de Dimensão “silver.dim_employee”

1. **employee_id (PK, STRING)**
   Identificador único do funcionário.

2. **employee_name (STRING)**
   Nome do funcionário.

3. **gender (STRING)**
   Gênero com o qual o funcionário se identifica.

4. **race (STRING)**
   Raça declarada pelo funcionário.

5. **marital_status (STRING)**
   Estado civil do funcionário.

6. **citizenship (STRING)**
   Situação de cidadania do funcionário.

7. **hispanic_latino (STRING)**
   Indicação se o funcionário se identifica como Hispânico/Latino (Sim ou Não).

C. Tabela de Dimensão “silver.dim_department”

1. **department_id (PK, STRING)**
   Identificador único do departamento.

2. **department_name (STRING)**
   Nome do departamento.

D. Tabela de Dimensão “silver.dim_position”

1. **position_id (PK, STRING)**
   Identificador único do cargo.

2. **position_name (STRING)**
   Nome do cargo ocupado pelo funcionário.

E. Tabela de Dimensão “silver.dim_manager”

1. **manager_id (PK, STRING)**
   Identificador único do gestor.

2. **manager_name (STRING)**
   Nome do gestor responsável.

3.3. Métricas

Com base no modelo de dados estruturado nas camadas Bronze, Silver e Gold, foi possível a definição e extração de métricas analíticas relevantes para a área de Recursos Humanos. Essas métricas foram consolidadas na camada **Gold**, que concentra dados agregados e prontos para análise gerencial, garantindo consistência, desempenho e facilidade de interpretação.

As métricas desenvolvidas permitem analisar aspectos fundamentais da gestão de pessoas, como **rotatividade**, **absenteísmo**, **desempenho**, **satisfação**, **diversidade** e **massa salarial**. A partir dessas informações, torna-se possível identificar padrões organizacionais, apoiar decisões estratégicas e fornecer subsídios para ações voltadas à melhoria do engajamento, da produtividade e da eficiência organizacional.

As métricas foram organizadas em tabelas temáticas, conforme descrito a seguir.

3.3.1. Métricas de Rotatividade (Turnover)

As métricas de rotatividade têm como objetivo analisar o comportamento de admissões e desligamentos dentro da organização, permitindo a avaliação da estabilidade do quadro de funcionários. Para isso, foram consideradas as seguintes análises:

- Quantidade total de funcionários por departamento;
- Total de funcionários desligados por departamento;
- Taxa de rotatividade calculada a partir da relação entre desligamentos e total de funcionários.

Essas métricas possibilitam identificar áreas com maior índice de desligamentos e apoiar ações de retenção de talentos.

3.3.2. Métricas de Absenteísmo

As métricas de absenteísmo permitem avaliar a frequência de ausências e atrasos dos colaboradores, fornecendo indícios sobre possíveis problemas de engajamento, clima organizacional ou sobrecarga de trabalho. As análises realizadas incluem:

- Total de ausências por departamento;
- Média de ausências por funcionário;
- Total de atrasos registrados nos últimos 30 dias;
- Comparação do nível de absenteísmo entre os departamentos da organização.

3.3.3. Métricas de Desempenho

As métricas de desempenho foram desenvolvidas para analisar a distribuição das avaliações de desempenho dos colaboradores, permitindo comparações entre áreas organizacionais. As análises contemplam:

- Distribuição das avaliações de desempenho por departamento;
- Quantidade de funcionários associada a cada nível de avaliação.

Essas métricas auxiliam na identificação de áreas com maior concentração de alto ou baixo desempenho.

3.3.4. Métricas de Satisfação dos Funcionários

As métricas de satisfação têm como objetivo avaliar a percepção dos colaboradores em relação à organização. A partir dos dados consolidados, foram realizadas as seguintes análises:

- Média do nível de satisfação dos funcionários por departamento;
- Identificação de departamentos com maior e menor índice médio de satisfação.

Essas informações são fundamentais para apoiar ações de melhoria do clima organizacional.

3.3.5. Métricas de Diversidade

As métricas de diversidade possibilitam a análise do perfil demográfico dos colaboradores, promovendo uma visão mais ampla sobre a composição da força de trabalho. As análises realizadas incluem:

- Distribuição de funcionários por raça, gênero e condição de cidadania;
- Comparação da diversidade entre os diferentes departamentos da organização.

Essas métricas apoiam iniciativas de inclusão e diversidade no ambiente corporativo.

3.3.6. Métricas de Massa Salarial

As métricas de massa salarial permitem analisar os custos relacionados à remuneração dos colaboradores, fornecendo uma visão consolidada por área organizacional. As análises contemplam:

- Massa salarial total por departamento;
- Salário médio por departamento;
- Salário mínimo e máximo por departamento;
- Quantidade total de funcionários por área.

Essas métricas auxiliam no controle orçamentário e na análise da distribuição salarial dentro da organização.

4.2. ETLs

A pipeline de dados deste projeto é composta por **três processos de ETL**, desenvolvidos em **notebooks Python**, sendo um para cada camada do Data Lake (Bronze, Silver e Gold). Esses processos têm como objetivo garantir a ingestão, tratamento, organização e disponibilização dos dados de forma estruturada, confiável e adequada para análise, conforme descrito a seguir.

4.2.1. ETL de Coleta de Dados (Camada Bronze)

O ETL responsável pela **coleta e ingestão dos dados na camada Bronze** foi implementado em um notebook Python. Esse processo realiza a leitura dos dados a partir de um **dataset público disponibilizado na plataforma Kaggle**, voltado à análise de recursos humanos (Human Resources Analytics).

A base de dados contém informações relacionadas a colaboradores, incluindo dados cadastrais, departamento, cargo, salário, desempenho, engajamento, ausências, atrasos, diversidade, além de informações sobre admissões e desligamentos. Esse conjunto de dados é amplamente utilizado em estudos de analytics e ciência de dados aplicados à gestão de pessoas.

Nesta etapa do pipeline, os dados são **armazenados em seu estado bruto**, sem aplicação de regras de negócio ou transformações complexas. São realizadas apenas ações mínimas necessárias para a ingestão, como:

- Leitura do arquivo original (formato .csv);
- Padronização básica de codificação e tipos de dados;
- Persistência dos dados na camada Bronze.

O objetivo principal da camada Bronze é **preservar a fidelidade da fonte original**, garantindo rastreabilidade e possibilitando reprocessamentos futuros, caso seja necessário. Dessa forma, essa camada atua como o ponto inicial e confiável do fluxo de dados do projeto.

4.2.2. ETL de análise e tratamento de dados (camada Silver)

O processo de ETL responsável pela análise, limpeza e transformação dos dados foi desenvolvido em um notebook Python utilizando PySpark e Spark SQL, totalmente documentado e executável no ambiente Databricks. O pipeline realiza a leitura dos dados da camada Bronze, aplica os tratamentos necessários para padronização e qualidade da informação e persiste os dados tratados na camada Silver.

Os dados foram carregados a partir da tabela bronze.rhg_dataset, que contém informações históricas de colaboradores, incluindo dados cadastrais, organizacionais, de desempenho e de desligamento. Inicialmente, foi conduzida uma análise exploratória por meio de consultas SQL, permitindo a verificação da estrutura do conjunto de dados, do volume total de registros e da distribuição das principais variáveis.

A análise de valores nulos indicou ausência de falhas relevantes nas colunas críticas para a análise, possibilitando o avanço para a etapa de tratamento sem a necessidade de imputações complexas. Em seguida, foi realizada a verificação de duplicidade com base na chave funcional id_funcionario, sendo identificados múltiplos registros para um mesmo colaborador. Essa característica foi interpretada como inerente à natureza histórica da base, representando diferentes eventos ao longo do tempo, como avaliações de desempenho, pesquisas de engajamento e atualizações cadastrais.

Na etapa de tratamento, foram aplicadas padronizações semânticas em campos categóricos relevantes, como departamento, cargo, avaliação de desempenho e nível de satisfação, por meio da remoção de espaços excedentes e da uniformização textual, com o objetivo de evitar inconsistências analíticas. Também foram realizados ajustes de tipos de dados, garantindo a correta representação de campos numéricos, datas e identificadores. As datas de contratação, desligamento e feedback foram mantidas no formato adequado, enquanto métricas quantitativas, como salário, ausências, atrasos e quantidade de projetos especiais, foram asseguradas como valores numéricos.

Após a aplicação dos tratamentos, os dados foram considerados consistentes e adequados para a modelagem analítica da camada Silver, viabilizando a criação de tabelas fato e dimensão voltadas à análise de retenção, desempenho e eficiência organizacional, sem a identificação de anomalias relevantes.

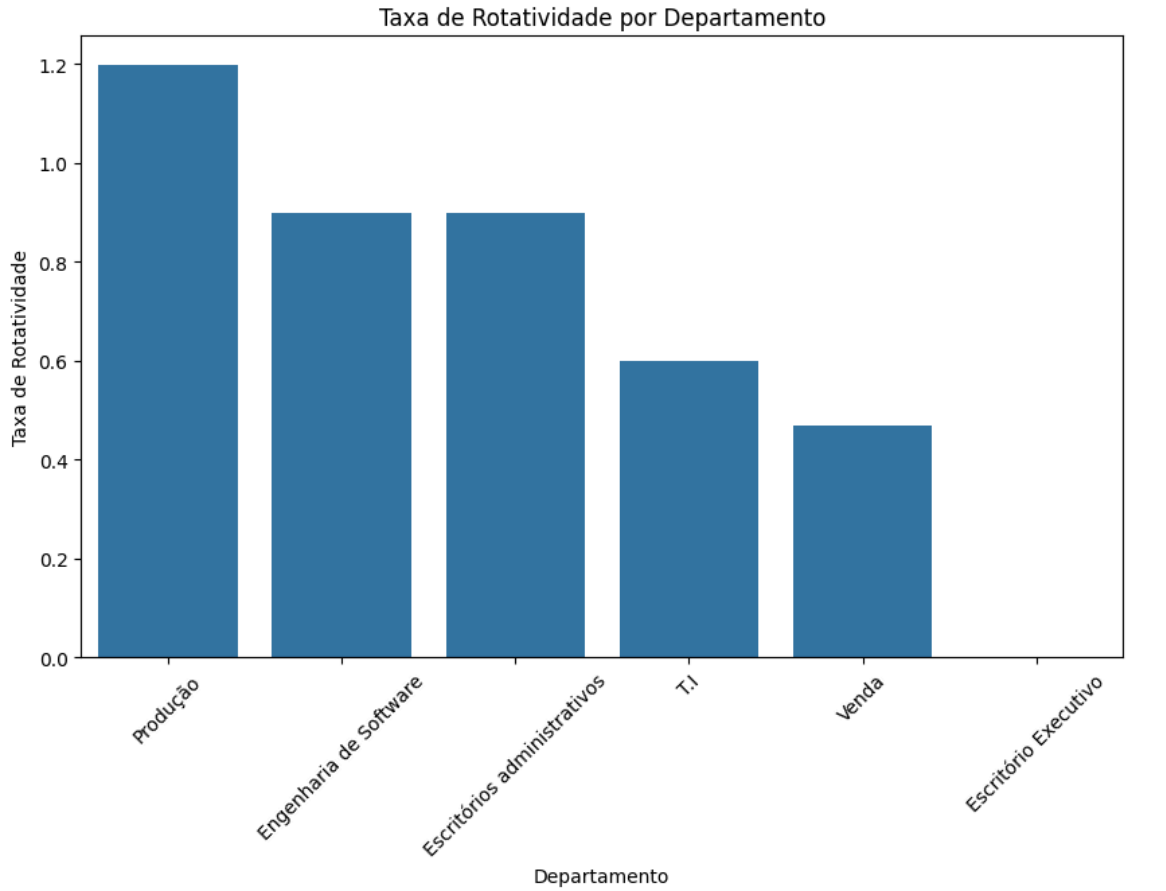
4.2.3. ETL de otimização de dados (métricas – camada gold)

O ETL responsável pelo enriquecimento dos dados e criação de métricas a partir deles está disponível no notebook. O notebook carrega os dados a partir da camada silver, enriquece os dados e cria métricas de negócio que são persistidos na camada gold

5. Autoavaliação:

A avaliação dos resultados é feita através das métricas e dos gráficos gerados a partir delas.

5.1 Análise da Rotatividade por Departamento



5.1.a Qualidade dos Dados

Para a análise de rotatividade (turnover), foram considerados os seguintes atributos do conjunto de dados: departamento, total de funcionários, total de desligados, taxa de rotatividade e tempo médio na empresa. A avaliação da qualidade dos dados indicou:

- Consistência: Todos os departamentos possuem números de funcionários e desligamentos coerentes. A taxa de rotatividade foi calculada corretamente, a partir da relação entre desligamentos e total de funcionários.

- Valores extremos: Alguns departamentos apresentaram valores atípicos que merecem atenção na interpretação:

    - O departamento “Escritório Executivo” possui apenas 1 funcionário, limitando a representatividade estatística.

    - O departamento “Produção” apresenta taxa de rotatividade acima de 1, indicando que houve mais desligamentos que o total de funcionários em determinado período. Esse valor não caracteriza erro, mas reflete alta rotatividade ou acumulação de desligamentos.

- Completude: Todos os atributos essenciais estão preenchidos, sem valores ausentes.

- Precisão temporal: O tempo médio na empresa, em dias, está coerente com a função ou senioridade dos colaboradores.

Conclusão sobre a qualidade dos dados:
O conjunto de dados está bem estruturado e pronto para análise. Pequenos pontos de atenção — como a amostra limitada do Escritório Executivo e a alta rotatividade em Produção — devem ser interpretados no contexto organizacional. Nenhuma limpeza adicional é necessária para responder à questão da rotatividade por departamento.

5.1.b Solução do Problema

A análise da rotatividade por departamento foi realizada a partir da tabela consolidada e do gráfico de barras que apresenta a taxa de rotatividade de cada área da organização.

Departamentos com maior rotatividade:

- Produção (1,1971): Apresenta a maior rotatividade, com mais desligamentos do que o total de funcionários em determinado período, indicando instabilidade significativa e risco de perda de talentos. Este cenário pode impactar diretamente a produtividade e a continuidade dos processos.

- Engenharia de Software e Escritórios Administrativos (0,9): Também apresentam taxas altas, embora inferiores à produção. Este resultado sugere necessidade de atenção em políticas de retenção e engajamento.

Departamentos com menor rotatividade:

- Vendas (0,4688) e T.I (0,6): Apresentam rotatividade moderada, indicando menor risco de perda de pessoal, mas ainda relevante.

- Escritório Executivo (0): Não apresenta desligamentos, mas o resultado deve ser interpretado com cautela, considerando a amostra de apenas 1 funcionário.

Implicações estratégicas:

- A alta rotatividade na Produção indica necessidade de investigação das causas de desligamento, como sobrecarga, clima organizacional, remuneração ou falta de oportunidades de desenvolvimento.

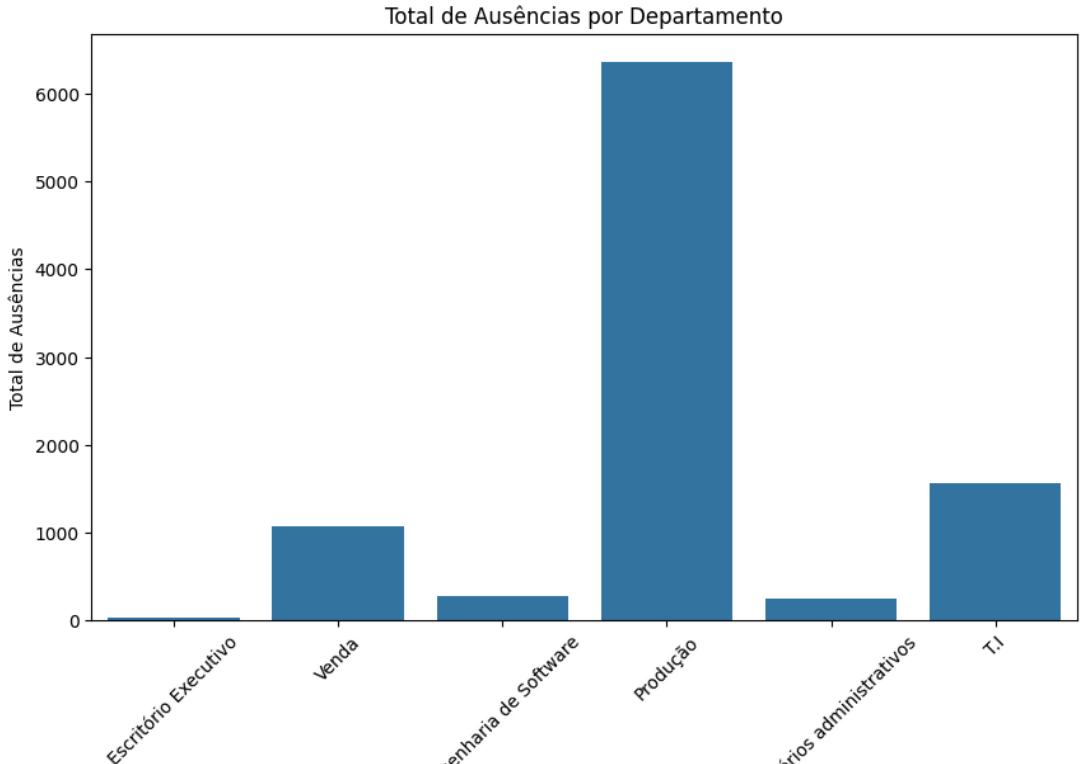
- Departamentos administrativos e de engenharia de software exigem atenção para desenvolver estratégias de retenção, incluindo benefícios, treinamentos e ações de engajamento.

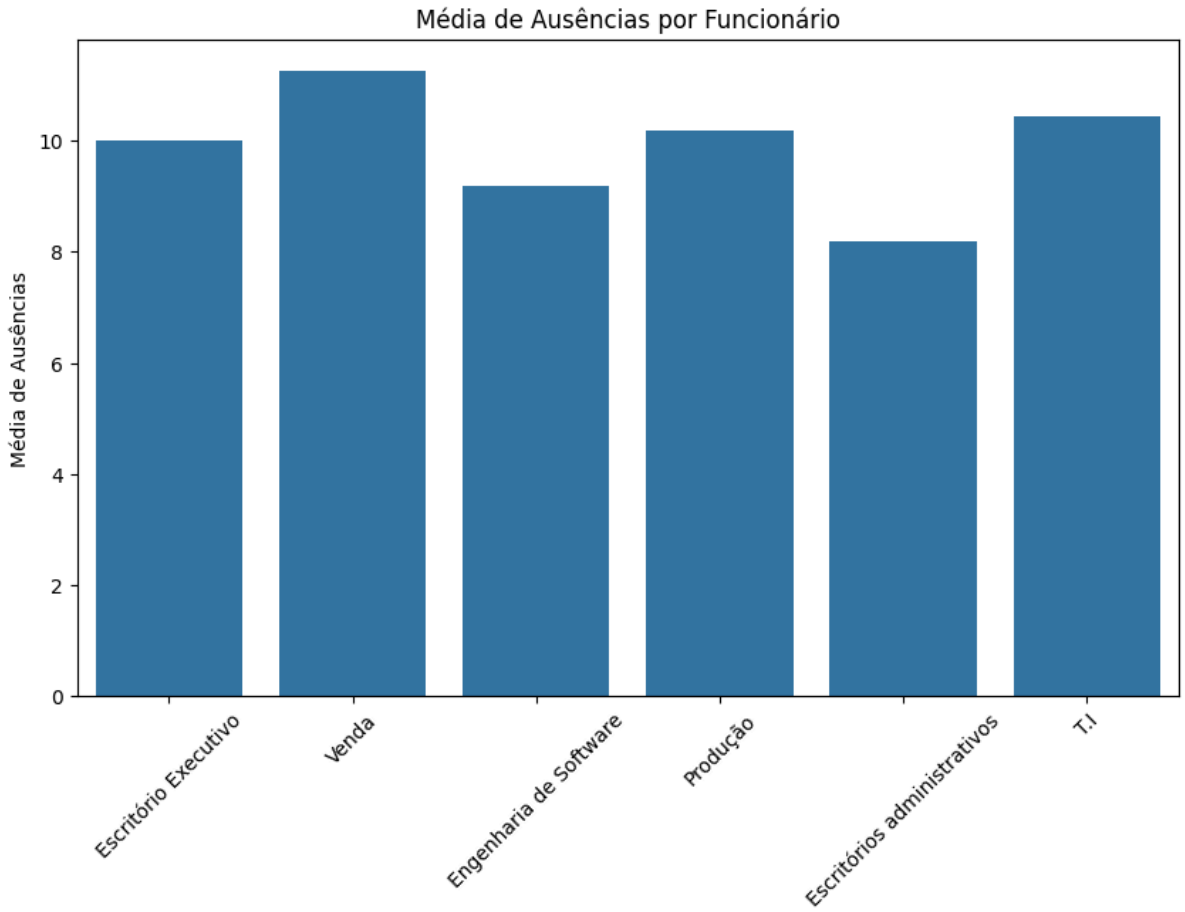
- Áreas com rotatividade baixa apresentam processos de gestão de pessoas mais eficazes ou menor pressão operacional, podendo servir de referência para práticas de RH em outros departamentos.

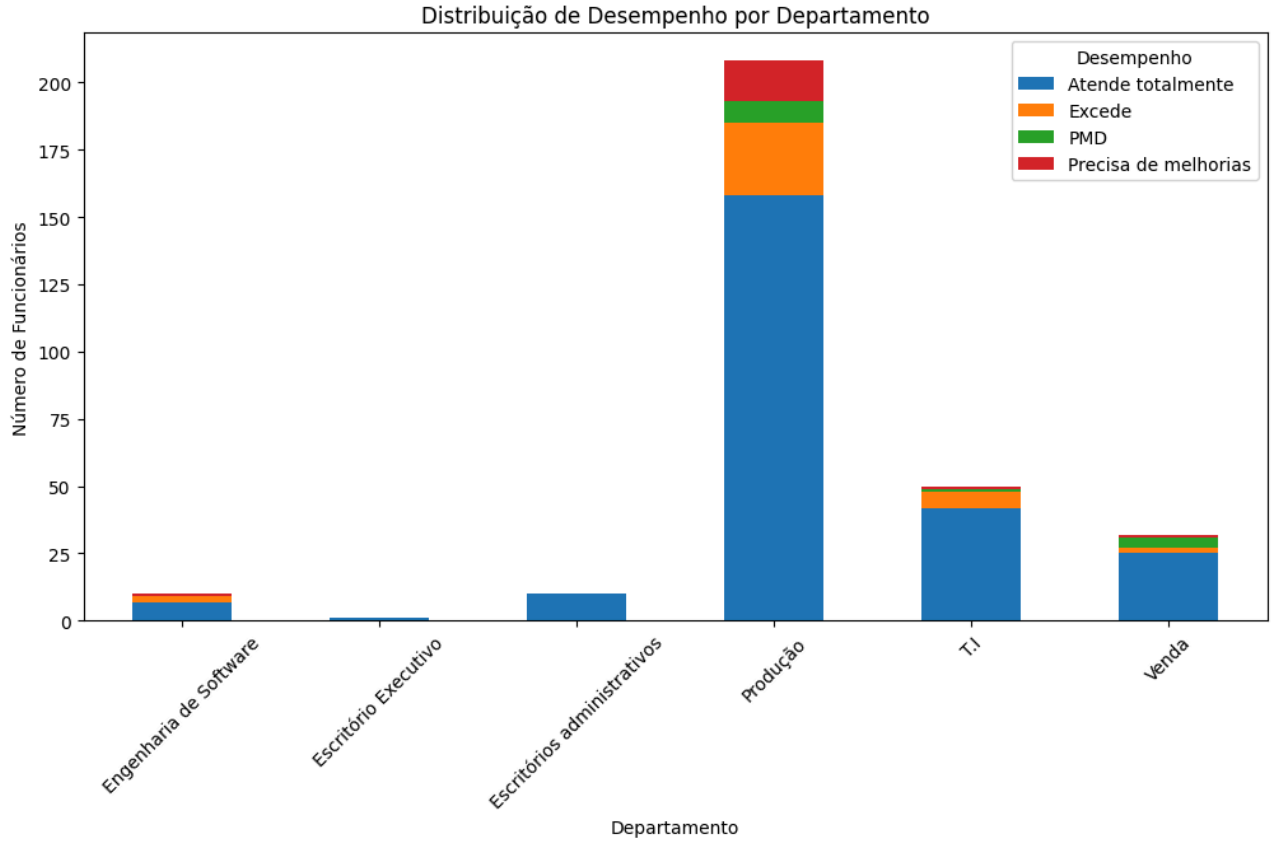
Conclusão geral:

A análise evidencia que a rotatividade não é uniforme entre os departamentos. A produção representa o maior risco de perda de colaboradores, seguida por Engenharia de Software e Escritórios Administrativos. A interpretação correta da taxa de rotatividade na composição dos departamentos, permite à organização direcionar ações estratégicas para retenção de talentos e melhoria do clima organizacional.

5.2 Análise do Absenteísmo por Departamento







5.2.a Qualidade dos Dados

Para a análise do absenteísmo, foram considerados os atributos: department, total_funcionarios,

total_ausencias, media_ausencias_por_funcionario, total_atrasos_ultimos_30_dias e performance_score (desempenho do funcionário).

- **Consistência:** Todos os departamentos possuem valores preenchidos para funcionários, ausências, atrasos e desempenho.

- **Cálculo das métricas:** A média de ausências por funcionário está coerente, derivada corretamente do total de ausências dividido pelo número de funcionários.

- **Valores extremos:**

    - Produção apresenta valores elevados de ausências (total 6.354) e atrasos (288), refletindo o contexto operacional.

    - Departamentos pequenos (Ex.: Escritório Executivo, T.I em algumas categorias, Venda com poucos funcionários) têm médias sensíveis a pequenas variações.

- **Completude:** Não há valores ausentes.

Conclusão sobre a qualidade dos dados:
O conjunto está completo e consistente. É importante interpretar resultados de departamentos com poucos funcionários com cautela, devido à amostra limitada.

5.2.b Solução do Problema

A análise do absenteísmo foi realizada considerando tanto o total de ausências por departamento quanto a média de ausências por funcionário, incorporando ainda a distribuição por performance.

Departamentos com maior absenteísmo:

- **Produção:**

    - Total de ausências: 6.354

    - Média por funcionário: 10,18

    - Total de atrasos: 288

    - Funcionários com desempenho `Precisa de melhorias` e `PMD` concentram maior parte dos atrasos (174 e 114, respectivamente), indicando que o absenteísmo está correlacionado com menor performance.

- **Vendas:**

    - Total de ausências: 1.080

- Média por funcionário: 11,25
- Total de atrasos: 66
- Colaboradores com desempenho `PMD` e `Precisa de melhorias` apresentam incidência significativa de ausências e atrasos, mostrando necessidade de acompanhamento individual e treinamento.

Departamentos com absenteísmo moderado ou baixo:

- **T.I:**
  - Média de ausências: 10,44
  - Atrasos: 21
  - Funcionários com desempenho `Precisa de melhorias` e `PMD` têm maior incidência de faltas, mas a maioria `Atende totalmente` ou `Excede` apresenta estabilidade.
- **Engenharia de Software:**
  - Média de ausências: 9,2
  - Atrasos: 12
  - Absenteísmo moderado, concentrado em funcionários com desempenho mais baixo.
- **Escritórios administrativos e Escritório Executivo:**
  - Ausências e atrasos baixos, refletindo menor pressão operacional e estabilidade nos processos administrativos.

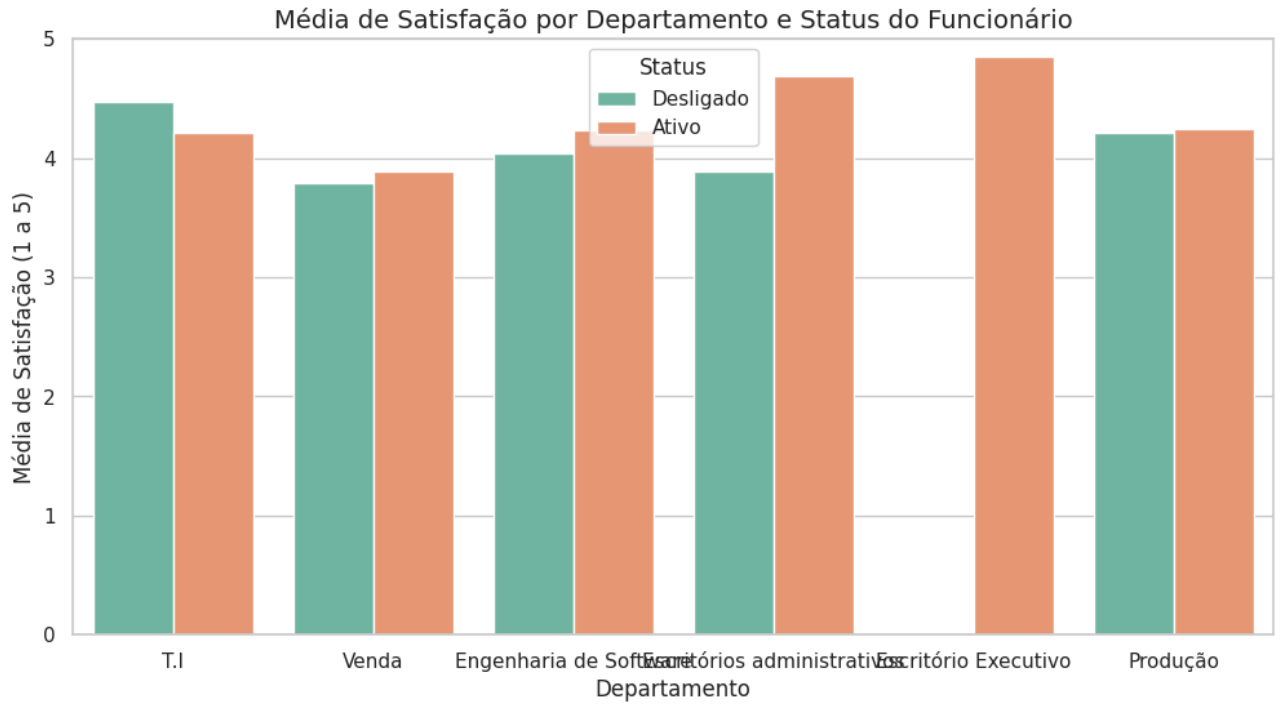
Interpretação estratégica:

- Existe uma clara **correlação entre menor desempenho e maior absenteísmo**, especialmente em departamentos operacionais e comerciais (Produção, Venda).
- A análise permite **priorizar ações de engajamento e acompanhamento individual** para colaboradores com desempenho crítico, reduzindo impactos na produtividade.
- Departamentos com desempenho consistente (`Atende totalmente` e `Excede`) apresentam absenteísmo mais controlado, servindo de referência para boas práticas de gestão de pessoas.

Conclusão geral:

O absenteísmo não é homogêneo entre departamentos e está relacionado ao desempenho dos colaboradores. Produção e Venda representam áreas críticas, com médias de ausências elevadas por funcionário e maior incidência entre colaboradores com performance baixa. Departamentos administrativos e técnicos apresentam menor absenteísmo e maior estabilidade. Esta análise permite à organização direcionar políticas de RH para reduzir ausências, melhorar engajamento e apoiar colaboradores com menor desempenho.

5.3 Análise de Satisfação dos Funcionários



5.3.a Qualidade dos Dados

Para a análise de satisfação, foram considerados os atributos: `department` (Departamento), `status_funcionario` (Ativo ou Desligado), `total_funcionarios` e `media_satisfacao`.

A avaliação da qualidade dos dados indicou:

- **Consistência:** Todos os departamentos possuem valores preenchidos, sem duplicidades ou registros ausentes.
- **Cálculo da média de satisfação:** A média foi corretamente normalizada para a escala de 1 a 5, garantindo coerência com os registros individuais de cada colaborador.
- **Valores extremos:** Alguns departamentos possuem número reduzido de funcionários em certas categorias (por exemplo, Escritório Executivo com 1 funcionário ativo), o que pode tornar a média mais sensível a variações individuais.
- **Completude:** Nenhum valor ausente foi identificado em nenhum dos atributos relevantes.

Conclusão sobre a qualidade dos dados:
O conjunto está completo, consistente e pronto para análise. Recomenda-se cautela na interpretação de departamentos com pequenas amostras devido à sensibilidade das médias.

5.3.b Solução do Problema

A análise considerou a média de satisfação dos funcionários por departamento, comparando colaboradores ativos e desligados.

Observações gerais por departamento:

- **Engenharia de Software:**
    - Ativos: 4,23
    - Desligados: 4,04
      A diferença indica que colaboradores mais satisfeitos tendem a permanecer, mostrando uma possível correlação entre satisfação e retenção.
- **T.I:**
    - Ativos: 4,21
    - Desligados: 4,47
      Curiosamente, desligados apresentam média ligeiramente maior que os ativos, sugerindo que outros fatores além da satisfação podem ter influenciado os desligamentos.
- **Produção:**
    - Ativos: 4,25
    - Desligados: 4,21
      Diferença pequena, indicando que a satisfação é relativamente homogênea e que o turnover pode estar relacionado a outros fatores, como condições de trabalho ou carga operacional.
- **Venda:**
    - Ativos: 3,89
    - Desligados: 3,79
      Satisfação mais baixa em geral, especialmente entre os desligados, sugerindo que melhorias no engajamento e clima podem impactar a retenção.
- **Escritórios Administrativos:**

    - Ativos: 4,69

    - Desligados: 3,89
      Diferença significativa, evidenciando forte correlação entre satisfação e permanência.

- **Escritório Executivo:**

    - Ativos: 4,85 (apenas 1 funcionário)
      Resultado alto, mas com amostra limitada, devendo ser interpretado com cautela.

Interpretação estratégica:

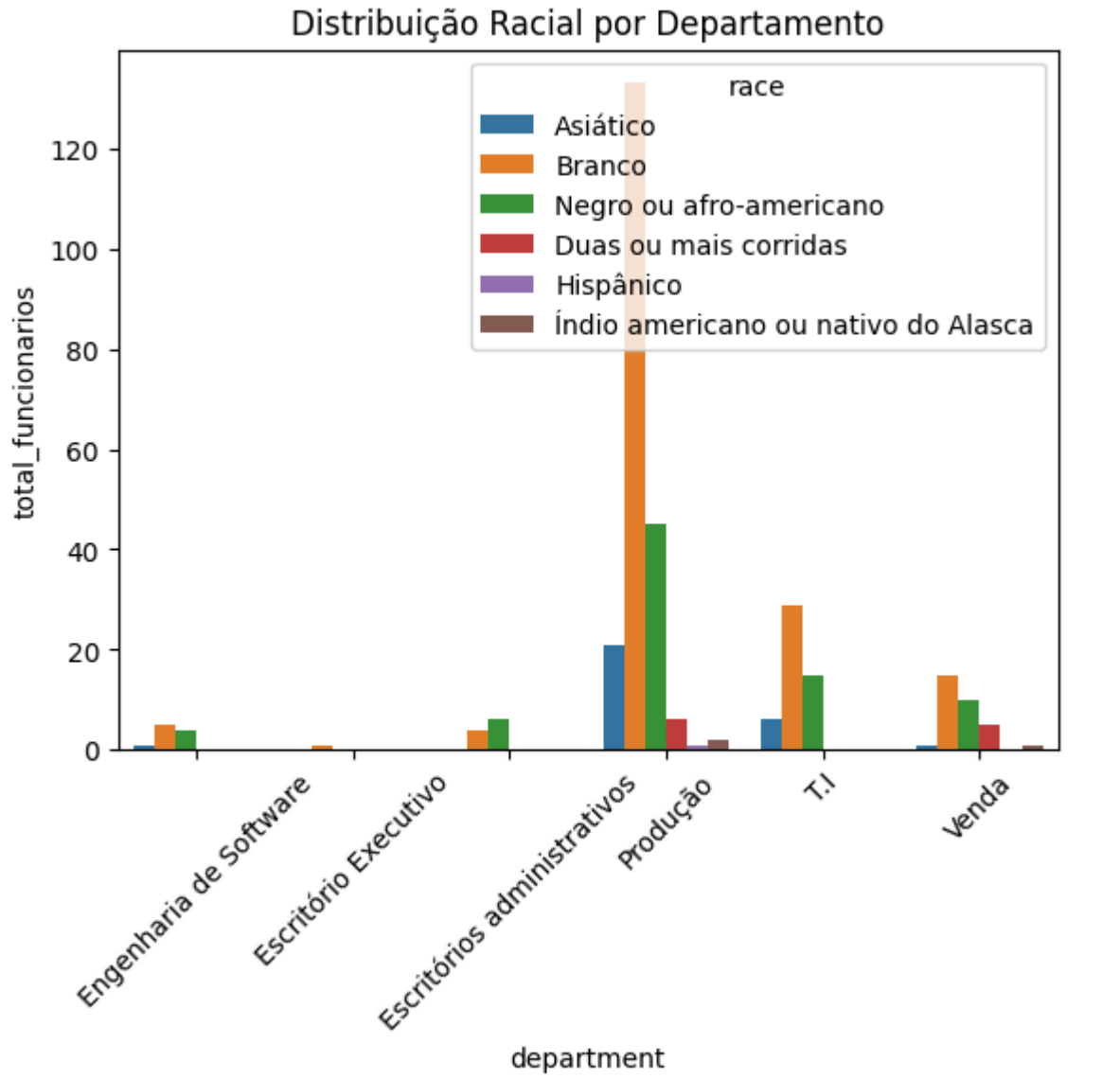
- Departamentos administrativos e de engenharia apresentam maior correlação entre satisfação e retenção.

- Departamentos operacionais (Produção e T.I) apresentam média de satisfação relativamente alta em ambos grupos, sugerindo que outros fatores podem influenciar o turnover.

- Área de Vendas apresenta menor satisfação geral, indicando necessidade de ações para engajamento e valorização dos colaboradores.

Conclusão geral:
A análise de satisfação evidencia que a percepção positiva dos colaboradores está relacionada à retenção em algumas áreas, principalmente administrativas e de engenharia. Entretanto, em departamentos operacionais, outros fatores — como condições de trabalho e carga operacional — também impactam o turnover. Esses resultados fornecem subsídios estratégicos para políticas de engajamento, clima organizacional e retenção adaptadas às particularidades de cada departamento.

5.4 Diversidade e Impacto na Retenção



5.4.a Qualidade dos Dados

Para a análise da diversidade, foram considerados os seguintes atributos: `department` (Departamento), `race` (Raça), `gender` (Gênero), `latino` (condição latino) e `total_funcionarios`.

A avaliação da qualidade dos dados indicou:

- **Consistência:** Todos os registros apresentam valores preenchidos e coerentes, sem duplicidades aparentes.
- **Completude:** Não há valores ausentes em nenhum dos atributos essenciais.
- **Uniformidade dos campos categóricos:** As categorias de raça, gênero e condição latino estão consistentes com padrões esperados, permitindo comparações entre departamentos e grupos.
- **Valores extremos:** Alguns departamentos têm pequenos grupos de funcionários em determinadas categorias (Ex.: Escritório Executivo com 1 funcionário), o que pode tornar médias e proporções mais sensíveis a alterações individuais.

Conclusão sobre a qualidade dos dados:

O conjunto de dados está adequado para análise, permitindo investigar padrões de diversidade e seu impacto potencial na retenção. Recomenda-se cautela na interpretação de categorias com pequenas amostras, evitando generalizações indevidas.

5.4.b Solução do Problema

A análise foi realizada com foco na distribuição de funcionários por raça, gênero e condição latino, associando essas informações à retenção (ativo vs desligado).

Observações gerais por departamento:

- **Produção:**
    - Maior diversidade racial, incluindo Brancos, Negros ou afro-americanos, Asiáticos, Hispanos e Duas ou mais corridas.
    - Predominância de funcionários brancos, mas presença significativa de outros grupos.
    - Departamentos com alta diversidade, como Produção, podem indicar a necessidade de políticas inclusivas e programas de engajamento para todos os grupos.
- **Engenharia de Software:**
    - Presença majoritária de Brancos e Negros ou afro-americanos.
    - Um equilíbrio razoável entre gêneros, porém predominância masculina.
- **T.I:**
    - Diversidade moderada, com funcionários Brancos, Negros ou afro-americanos e Asiáticos.
    - Alguns grupos pequenos de latinos e minorias raciais, sugerindo atenção à inclusão.
- **Vendas:**
    - Diversidade variada, mas predominância de Brancos e Negros ou afro-americanos.
    - Alguns grupos pequenos (Índio americano ou nativo do Alasca, Asiático) com baixo número de funcionários.
- **Escritórios Administrativos:**
    - Menor diversidade relativa, com predominância de Brancos e Negros ou afro-americanos.

    - Pequeno número de funcionários em algumas categorias.

- **Escritório Executivo:**

    - Apenas 1 funcionário, Branco, latino, o que limita qualquer análise de diversidade ou impacto na retenção.

Interpretação estratégica:

- Departamentos com maior diversidade, como Produção e T.I, não apresentam necessariamente maior rotatividade, sugerindo que políticas inclusivas e de engajamento são importantes para manter talentos em todos os grupos.

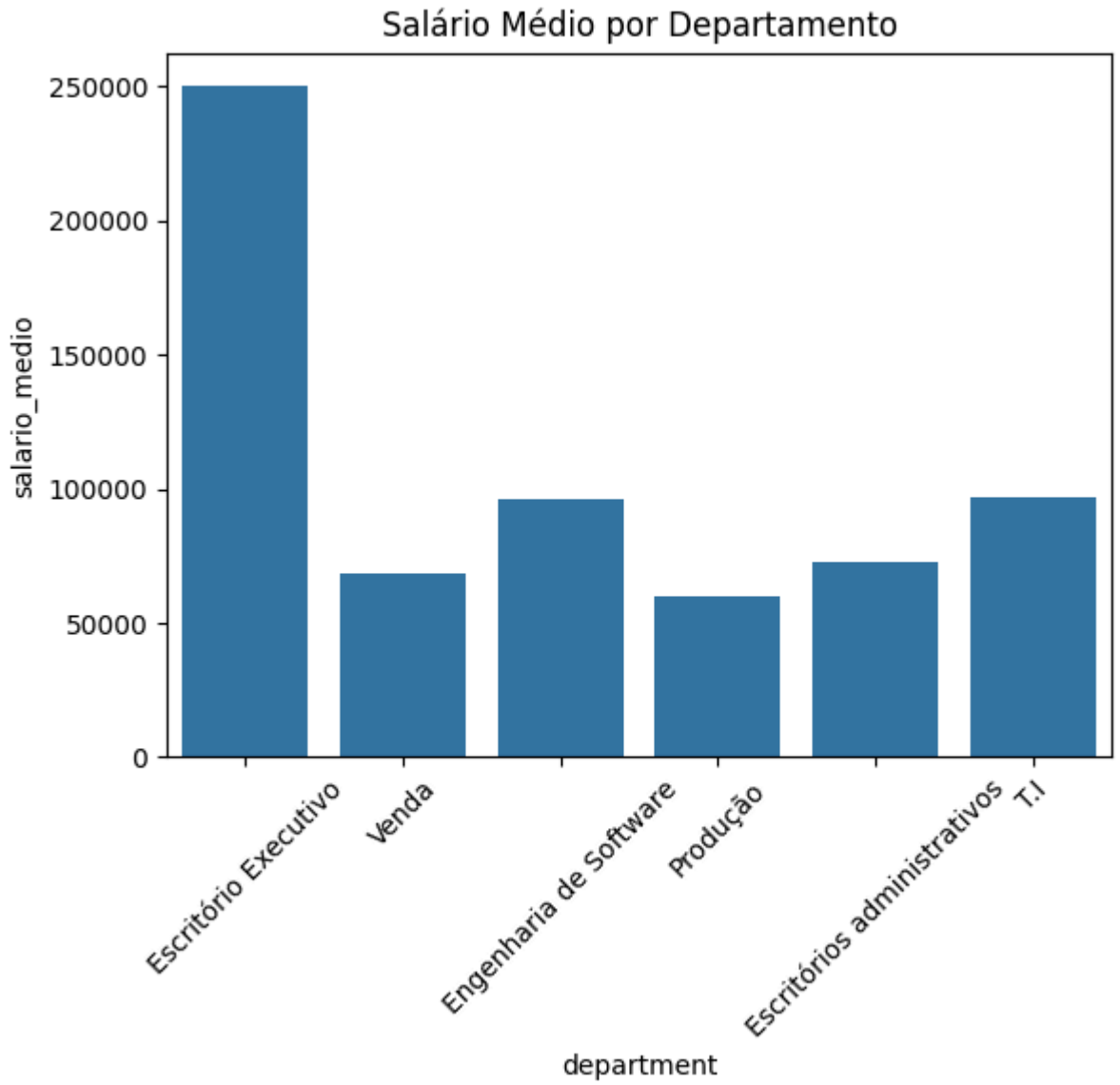
- A análise indica que diversidade racial, de gênero e condição latino não é, isoladamente, um fator determinante da rotatividade, mas deve ser considerada em conjunto com satisfação, desempenho e condições de trabalho.

- Pequenos grupos em qualquer categoria racial ou de gênero demandam atenção para evitar isolamento ou desmotivação, garantindo que todos se sintam incluídos e representados.

Conclusão geral:

A organização apresenta níveis variados de diversidade entre os departamentos. Produção e T.I se destacam pela diversidade racial e de gênero, enquanto áreas administrativas e executivas têm menor variação. A retenção não parece ser impactada diretamente pela diversidade, mas políticas de inclusão e engajamento são essenciais para manter o bem-estar e motivação dos colaboradores, especialmente em departamentos com menor representatividade de minorias.

5.5 Análise da Massa Salarial e Média Salarial por Departamento



5.5.a Qualidade dos Dados

Para a análise da massa salarial e da média salarial, foram considerados os seguintes atributos: `department` (Departamento), `total_funcionarios`, `massa_salarial_total`, `salario_medio`, `salario_minimo` e `salario_maximo`.

A avaliação da qualidade dos dados indicou:

- **Consistência:** Todos os departamentos possuem valores preenchidos para todos os atributos, sem registros ausentes.
- **Cálculo da média salarial:** O salário médio está coerente com a massa salarial total dividida pelo total de funcionários.
- **Valores extremos:**
    - O Escritório Executivo possui apenas 1 funcionário, com salário fixo de 250.000, o que gera média, mínimo e máximo iguais.

    - Produção apresenta salário máximo significativamente maior que a média, refletindo a presença de cargos de maior senioridade ou funções estratégicas.

- **Completude:** Não foram identificados valores ausentes.

Conclusão sobre a qualidade dos dados:

O conjunto de dados está completo, consistente e adequado para análise. É importante interpretar os departamentos com pequenas amostras (como Escritório Executivo) com cautela, evitando conclusões generalizadas.

5.5.b Solução do Problema

A análise da massa salarial e dos salários médios permite compreender a distribuição de custos com pessoal por departamento e identificar possíveis diferenças salariais significativas entre áreas e cargos.

Observações por departamento:

- **Produção:**
    - Massa salarial total: 37.419.900
    - Salário médio: 59.967,79
    - Salário mínimo: 45.046
    - Salário máximo: 170.500
    - Apesar da maior quantidade de funcionários (208), o salário médio é menor comparado a áreas técnicas, mas o salário máximo elevado indica cargos estratégicos ou especializados.
- **T.I:**
    - Massa salarial: 14.559.696
    - Salário médio: 97.064,64
    - Salário mínimo: 50.178
    - Salário máximo: 220.450
    - Departamento técnico com salários médios elevados e maior variação entre mínimo e máximo, refletindo diferenciação por cargo e senioridade.
- **Venda:**

    - Massa salarial: 6.593.670
    - Salário médio: 68.684,06
    - Salário mínimo: 55.875
    - Salário máximo: 180.000
    - Média salarial intermediária, mas com presença de cargos comerciais estratégicos com salários altos.

- **Engenharia de Software:**
    - Massa salarial: 2.884.563
    - Salário médio: 96.152,1
    - Salário mínimo: 77.692
    - Salário máximo: 108.987
    - Pequeno departamento com salários médios altos, refletindo a especialização e valor estratégico dos colaboradores.

- **Escritórios Administrativos:**
    - Massa salarial: 2.188.470
    - Salário médio: 72.949
    - Salário mínimo: 49.920
    - Salário máximo: 106.367
    - Média salarial razoável, adequada à função administrativa, com variação moderada.

- **Escritório Executivo:**
    - Massa salarial: 750.000
    - Salário médio: 250.000
    - Salário mínimo e máximo: 250.000
    - Departamento com um único funcionário, refletindo alta remuneração executiva, sem variação salarial.

Interpretação estratégica:

- Departamentos técnicos (T.I e Engenharia de Software) apresentam salários médios mais elevados, refletindo a especialização e o valor estratégico dos colaboradores.

- Produção, com maior número de funcionários, concentra custos salariais elevados em massa total, mas salário médio inferior, indicando maior proporção de funções operacionais.

- Departamentos administrativos e comerciais apresentam médias intermediárias, compatíveis com funções de suporte e vendas, com alguns cargos estratégicos elevando o teto salarial.

- Escritório Executivo concentra alto salário individual, típico de cargos de liderança sênior.

Conclusão geral:

A análise da massa e média salarial evidencia diferenças claras entre departamentos, refletindo a natureza das funções e o valor estratégico de cada área. A informação permite à organização tomar decisões sobre políticas salariais, planejamento de orçamento de pessoal e estratégias de retenção e valorização de talentos em áreas críticas.

6. Conclusão Final

O presente estudo permitiu analisar de forma integrada os principais aspectos da gestão de pessoas na organização, utilizando dados de RH consolidados em uma pipeline de dados estruturada. As análises demonstraram que a rotatividade, o absenteísmo, o desempenho, a satisfação, a diversidade e a distribuição salarial variam significativamente entre departamentos, refletindo diferenças de função, carga operacional e condições de trabalho.

A alta rotatividade e absenteísmo observados em Produção e Venda indicam a necessidade de políticas de engajamento, acompanhamento individual e estratégias de retenção, enquanto áreas administrativas e técnicas apresentam maior estabilidade e satisfação, servindo de referência para boas práticas. A diversidade, embora não seja determinante isoladamente para a retenção, evidencia a importância de políticas inclusivas e de engajamento para manter talentos em todos os grupos. A análise salarial mostrou que departamentos estratégicos e técnicos apresentam maior remuneração média, enquanto áreas operacionais concentram maior massa salarial devido ao volume de funcionários.

A criação da pipeline de dados, estruturando informações desde a ingestão (camada Bronze) até a geração de métricas analíticas (camada Gold), foi fundamental para garantir a qualidade, consistência e rastreabilidade dos dados. Este processo permitiu transformar dados brutos em insights estratégicos, possibilitando tomadas de decisão mais embasadas, ágeis e confiáveis.

Em síntese, a integração entre análise de dados e construção de pipelines oferece uma visão completa do capital humano da organização, destacando áreas críticas, apoiando políticas de retenção, engajamento e remuneração, e demonstrando o valor da Ciência de Dados aplicada à gestão de pessoas.