UNIVERSIDADE LUSÓFONA DE HUMANIDADES E TECNOLOGIA
CUL LISBOA

Thiago Gustavo Vieira de Paiva

RELATÓRIO DE ATIVIDADES PARA A EMPRESA DE CONTABILIDADE

Lisboa
2023

THIAGO GUSTAVO VIEIRA DE PAIVA

RELATÓRIO DE ATIVIDADES PARA A EMPRESA DE CONTABILIDADE

Relatório apresentado ao curso de Mestrado em Ciência de Dados da Universidade Lusófona de Humanidades e tecnologia, como parte necessária de avaliação da Visualização de Informação.

Lisboa, 03 de Junho de 2023

Resumo

O presente relatório visa o detalhamento de atividades desenvolvidas no âmbito do Projecto da Empresa de Contabilidade fictícia envolvendo todo o processo de extração-transformação-carregamento dos dados.

Ao longo deste relatório, serão abordados os procedimentos adotados quanto a tarefas relacionadas ao processamento e tratamento prévio dos dados fornecidos, medidas adotadas e eventuais transformações realizadas e, por fim, exibição de conclusões a partir da Análise Exploratória dos Dados identificando padrões nos dados.

Palavras-chave: extração-transformação-carregamento, análise exploratória dos dados, padrões nos dados.

Abstract

The present report aims to provide a detailed account of activities carried out within the scope of the fictional Accounting Company Project, involving the entire data extraction-transformation-loading process.

Throughout this report, the procedures adopted for tasks related to data processing and pre-processing, the measures taken, and any transformations performed will be discussed. Finally, conclusions will be presented based on Exploratory Data Analysis, identifying patterns in the data.

Keywords:extract-transform-load, exploratory data analysis, data patterns.

Sumário

1 Introdução

A Empresa Fictícia de Contabilidade doravante tratada também por *EFC*, possui dados descentralizados, distribuídos por planilhas no formato *.xlsx*.

São carecterísticas destes dados o fato de ser necessária uma consolidação e validação destes no sentido de garantir consistência segura para sua análise e identificação de padrões.

O objetivo é, portanto, analisar e identificar padrões nos dados.

A partir dos dados fornecidos efetuar uma análise inicial a fim de perceber o quão bem estruturados estão tais dados.

Tratar tais dados a fim de garantir solidez o bastante para que sejam analisados mais profundamente e, por fim, gerar conclusões e extrair valor dos dados fazendo uso de visualizações interativas que facilitem o entendimento da produtividade de tal empresa.

2 Objetivos

2.1 Gerais

Este relatório tem como principal objetivo analisar e identificar padrões e extraindo valor dos dados.

2.2 Específicos

Como parte do processo de análise realizar a consolidação e validação dos dados facilitando análises futuras. Este trabalho pretende então:

1. Extrair os dados fornecidos pela fonte primária Produtividade.xlsx;
2. Tratar os dados centralizando-os num banco de dados relacional normalizado.
3. Gerar um novo *dataset* extraído diretamente da base de dados garantindo consequentemente a consistência dos dados.
4. Realizar a Análise Exploratória dos Dados a fim de identificar caracteríticas extraindo informações, valor e identificando padrões.
5. Produzir *Dashboard* para visualizações necessárias.

3 Metodologia

A metodologia utilizada neste projeto de ciência de dados seguiu as etapas descritas a seguir:

1. Extração de Dados:
    - Inicialmente, os dados foram extraídos do dataset fornecido, que consistia em uma planilha Excel chamada "Produtividade.xlsx".
    - A planilha continha três abas: "RegistoTempos", "Tarefas", "Clientes"e "Colaboradores".
    - Na aba "RegistoTempos", foram extraídas as colunas `Data`, `Hora Início`, `Hora Fim`, `Horas`, `Colaborador`, `Tarefa` e `Cliente`.
    - Na aba "Tarefas", foram obtidos os dados das colunas `COD_Tarefa`, `Tarefa` e `Descrição`.
    - Na aba "Clientes", foram extraídas as informações de `Cod_Cliente`, `Cliente`, `Morada`, `Distrito`, `Colaborador`, `Valor_Hora`, `Horas Mensais Orçamentadas` e `Satisfação Cliente (1-10)`.
    - Por fim, na aba "Colaboradores", foram extraídos os dados de `Nome`, `Salário Base`, `Horas Semana`, `Equipa`, `Idade` e `Qualificações` dos colaboradores.
2. Modelagem de Dados:
    - Com base na análise inicial dos dados fornecidos, foi decidido criar uma base de dados relacional para garantir a integridade e consistência dos relacionamentos entre os dados.
    - A base de dados MariaDB (https://mariadb.org/) foi escolhida devido à sua natureza *open source* e capacidade de suportar uma estrutura relacional.
    - O modelo do banco de dados utilizado foi o seguinte:
        - Tabela Clientes (**id_cliente**, `nome`, `morada`, `distrito`, `valor_hora`, `horas_mensais_o` `satisfacao`, id_colaborador).
        - Tabela Colaboradores (**id_colaborador**, `nome`, `salario_base`, `horas_semanais`, `idade`, `qualificacao`, id_equipa).
        - Tabela Equipas (**id_equipa**, `nome`).
        - Tabela Tarefas (**id_tarefa**, `nome`, `descricao`).
        - Tabela Rel_Clientes_Colaboradores_Tarefas (**id_rel**, **id_cliente**, **id_colaborador**, **id_tarefa**, `data`, `hora_inicio`, `hora_fim`, `horas`).

3. Carregamento dos Dados:

    - Os dados extraídos foram carregados nas respectivas tabelas do banco de dados MariaDB, seguindo a estrutura definida.
    - Foram estabelecidas as relações entre as tabelas por meio das chaves primárias e chaves estrangeiras, garantindo a integridade referencial.
    - *SQL queries* foram criadas para a inserção de dados em conformidade com tal modelo e estas foram obtidas diretamente a partir de *Pandas.DataFrame* utilizados para montar automaticamente as *insert queries* para cada tabela.
    - Estas *queries* foram salvas no diretório /**sql** para fins de verificação.
    - As *queries* foram executadas fazendo uso de um objeto de conexão direta com o banco de dados, criado e importado de **database_connections.py** com dados para sua conexão com a base de dados em **config/db_config.json**. Desta forma, todas as *queries* puderam ser inseridas e consultadas a partir do *Jupyter Notebook* nomeado **Reverse_Engineering.ipynb** localizado em /**src**
    - A construção de um dataframe consilidado após a normalização das tabelas permitiu mesclar (*merge*) os dados com efetividade e consistência de modo que estes dados foram salvos em **data/dataset.csv** e seu *schema de dtypes* em **config/dtypes.json**.

4. Análise de Dados:

    - O *loading dos dados* diretamente de *data/dataset.csv* permitiu lidar com os dados de maneira mais simples a efetuar análises e gerar visualizações com auxílio da biblioteca gŕafica **plotly** (https://plotly.com/).
    - Conclusões foram obtidas em considerações aos dados.
    - Um dashboard com algumas visualizações relevantes foi criado em **src/dashboard.py**.
    - O processo de análise encerrou-se com a construção deste Relatório.

4 Análise Exploratória de Dados

Esta fase iniciou com o carregamento dos dados a partir de **data/dataset.csv** que contém os dados da Empresa Fictícia de Contabilidade - EFC de maneira centralizada e consistente apresentando as seguintes características:

- Um total de **52416 linhas** e **23 colunas**.
- Quanto ao tipo de *features*
  - Date Features: 3
  - Categorical Features: 5
  - Numerical Features: 15
- Colunas inteiramente com valores nulos: **morada**, **distrito** e **descricao** que foram posteriormente removidas diminuindo a dimensão de colunas para 20.
- Nenhum registo duplicado.

Tendo em vista estas informações, foi realizada a análise quanto aos Colaboradores de acordo com o seguinte gráfico:

Figura 1 – Registos por Colaboradores

O número de registos por Clientes foi então gerado a fim de compreender como se dá a frequência de registos por parte destes o que pode ser visto com a imagem a seguir.

Em seguida, o atendimento de Clientes por parte dos Colaboradores foi relacionado conforme o gráfico que segue:

Figura 2 – Registos por Clientes

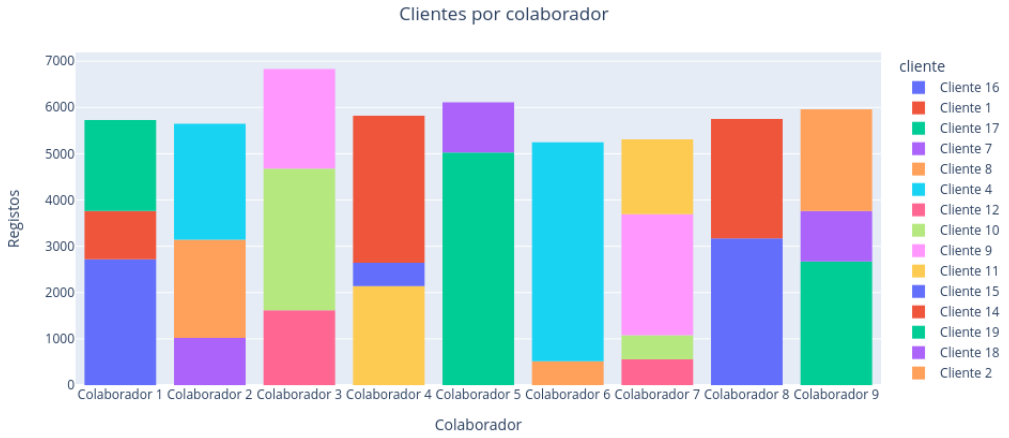

Figura 3 – Clientes por Colaborador

- Estas visualizações permitiram concluir que:
  - Existem 9 Colaboradores nesta empresa.
  - A maioria dos colaboradores atende a 3 clientes.
  - **Colaborador 3** teve a maior quantidade de atendimento seguido pelo **Colaborador 5** e **Colaborador 9**.
  - O **Colaborador 7** foi aquele com maior diversidade de clientes ao atender a 4 clientes.
  - **Colaborador 5** e **Colaborador 6** tiveram a menor diversidade de clientes atendendo cada um a apenas 2 clientes, o que pode apresentar-se como fator positivo em razão de lidarem com diversos clientes.

Dando prosseguimento à Análise Exploratória, as Tarefas foram verificadas a fim de ser determinado quais era as mais frequentemente realizadas segundo o gráfico a seguir.

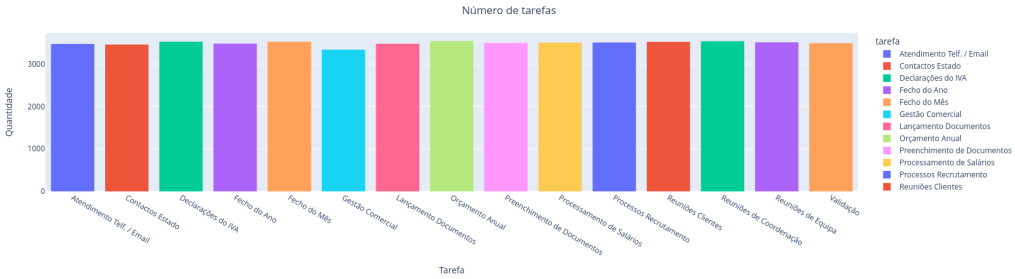

Figura 4 – Número de Tarefas

Quanto a realização destas pelos colaboradores, podemos verificar abaixo.

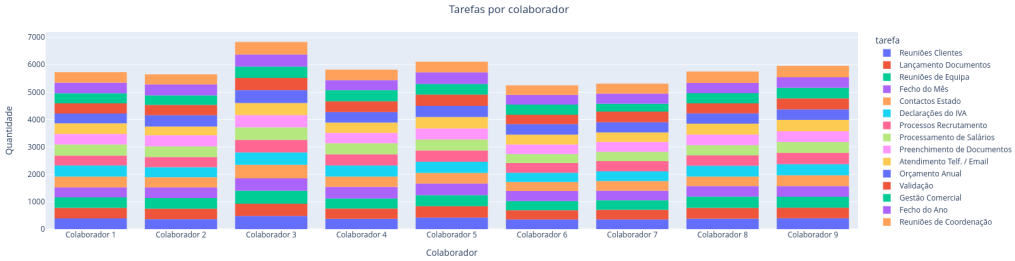

Figura 5 – Tarefas por Colaborador

- Destas visualizações conclui-se que:
  - **Orçamento Anual**, **Reuniões de Coordenação** e **Fecho do Mês** são as tarefas mais realizadas.
  - Há uma distribuição equilibrada tanto em termos de realizações quanto à distribuição entre os colaboradores.

A análise dos salários e qualificação dos colaboradores foi feita e tomando como ponto de partida o gráfico:

Figura 6 – Salário base por Colaborador

Nota-se que o salário base varia com o valor mínimo de 850 a 1100 unidade monetária.

Quanto a qualificação de cada Colaborador, com interesse em suas remunerações média, temos:

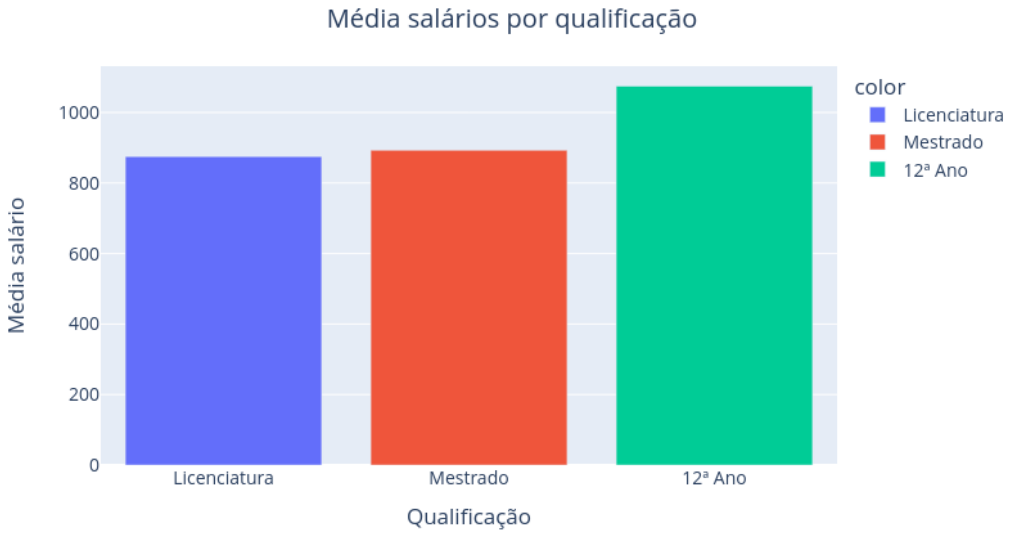

Figura 7 – Média de salários por qualificação

O que, apesar de trazer certa surpresa dada a qualificação, quando confrontada diante do maior peso em folha salarial, constata-se que colaboradores mais qualificados possuem maior valor acumulado conforme verifica-se logo adiante.

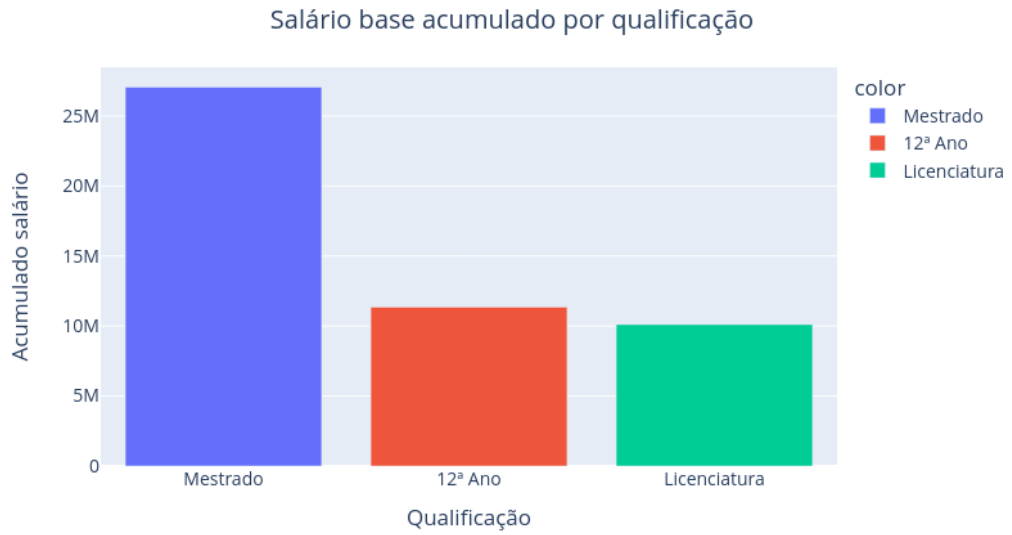

Figura 8 – Salário acumulado por Colaborador em vista a sua qualificação

Os fatores que contribuem para maior média entre aqueles colaboradores com menor qualificação pode se dar dada sua antiguidade ou outros fatores não revelados por estes dados.

Em relação à Equipas, os dados revelaram que a SEDE possui mais colaboradores contando com 3 destes e as demais equipas em PORTIMÃO, FARO E LAGOAS contam com 2 colaboradores cada o que explica bem o motivo da quantidade de registos maior ter sido atendida por colaboradores da SEDE conforme a imagem a seguir.

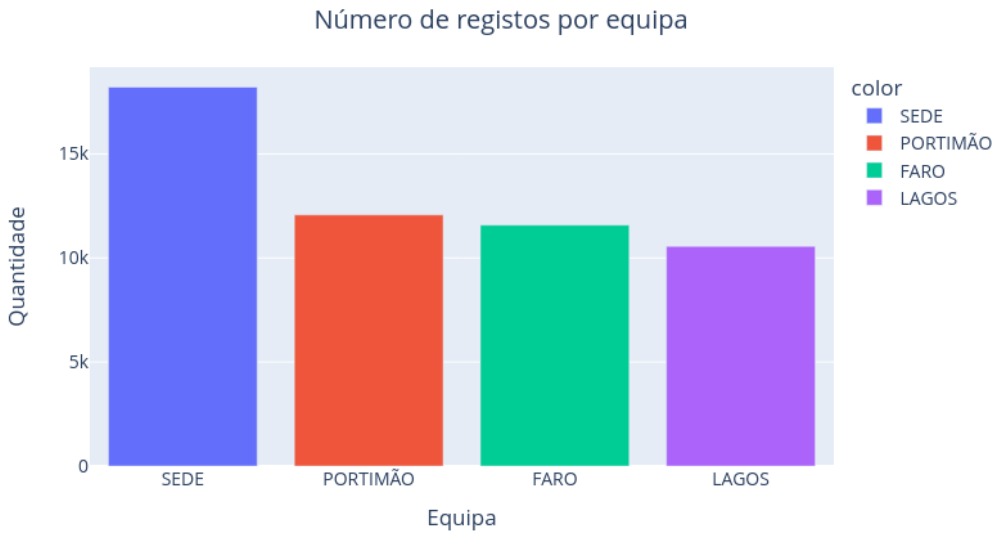

Figura 9 – Registos por Equipa

A análise quanto a temporalidade dos registos de tarefas foi realizada, mas não foi identificado grande fator determinante a não ser que:

- A série temporal inicia-se em 2021-01-02 00:00:00 e termina em 2021-12-31 00:00:00.
- A hora de início das tarefas se deu com maior frequência em 00:00:00 o que revela que a mesma pode não ter sido registada durante o atendimento aos clientes.
- Atendimentos se deram de maneira equilibrada ao longo dos dias de modo a se parecer com ruído branco como visto abaixo.

A análise exploratória foi encerrada com a análise temporal, mas o dashboard ainda revela características que não foram levadas em consideração no âmbito da EDA.

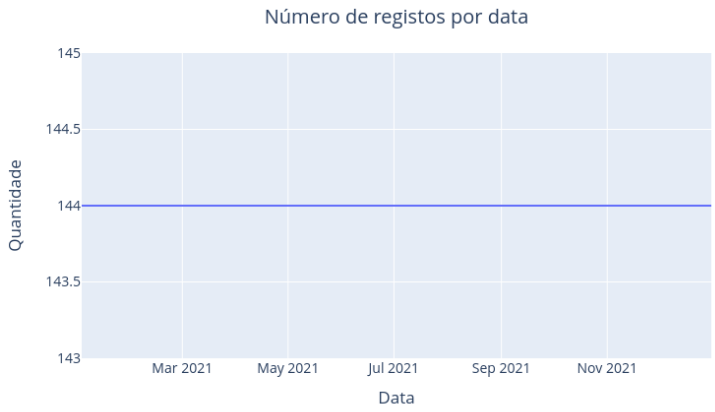

Figura 10 – Registos por Data

5 Considerações Finais

A proposta principal do Projecto da Empresa Fictícia de Contabilidade envolvia analisar os dados fornecidos e identificar padrões.

Durante o desenvolvimento do projecto, as necessidades inerentes de extração, tratamento e carregamento dos dados foram atingidas por meio das soluções técnicas e metodologia científica a ponto de extrair valor dos dados em todas as fases.

Entende-se que tanto os objetivos gerais quanto os específicos propostos puderam levar a um melhor entendimento da empresa levantando conclusões que foram verificadas tanto na Análise Exploratória dos Dados, quanto em fases anteriores e posteriores a esta visto que, a construção de dashboard permitiu ainda constatar visualizações e linhas de análise não previstas durante a EDA de modo a complementar as conclusões obtidas.