UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

JÚLIA DEL PINO RITTMANN

Análise de três dimensões de qualidade de dados em tabelas de compras de um ambiente de Data Warehouse

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Engenharia da Computação

Orientador: Prof[a]. Dra. Renata de Matos Galante

Porto Alegre
2023

AGRADECIMENTOS

Aos meus pais, Daisy Lopes Del Pino e Ricardo Rittmann, que sempre colocaram o estudo meu e de meus irmãos como prioridade dos seus trabalhos, o que possibilitou que tivéssemoss acesso às melhores oportunidades de educação. Mais ainda, dedico a terem nos criado para sermos pessoas que sempre buscam estudar e se profissionalizar, pois acredito que isto tenha sido fundamental para me tornar a profissional que sou hoje. Agradeço pelo apoio incondicional que possibilitou esta conquista.

Aos meus irmãos, Gabriel Del Pino Rittmann e Rodrigo Del Pino Rittmann, que sempre me apoiaram e criaram momentos divertidos de descontração, os quais tornaram esta experiência mais alegre.

Aos meus amigos, por tornarem a graduação mais divertida e prazeroza. Também por me ouvirem em momentos difíceis e me apoiarem a seguir confiante nesta etapa.

À Professora Renata de Matos Galante pela orientação e pelos incentivos, bem como a confiança depositada para que a escolha deste trabalho fosse possível.

RESUMO

Nos últimos anos, o aumento exponencial na criação de dados em todo o mundo tem despertado a atenção das grandes empresas para a importância de cuidados adequados com esse valioso recurso. Um dos critérios essenciais para obter projeções e análises precisas dos dados corporativos é a qualidade desses dados, uma vez que é fundamental garantir a integridade dos dados utilizados em qualquer estudo. Embora seja um tópico relativamente novo no mercado, é crucial considerar as formas de implementar a qualidade dos dados em empresas de pequeno e grande porte. Este estudo tem como objetivo desenvolver um código em Python para analisar três das seis dimensões reconhecidas da qualidade dos dados: completude, conformidade e precisão. Essa análise é realizada em tabelas de um *Data Warehouse* de uma empresa brasileira do setor varejista. Essas análises são fundamentais para a correta varredura e interpretação das bases de dados em questão, permitindo a criação de Indicadores-Chave de Desempenho (*KPIs*) mais confiáveis. O trabalho busca abrir caminho para o aprimoramento das práticas relacionadas à qualidade de dados, demonstrar a eficácia da aplicação da qualidade dos dados e evidenciar o impacto que a atenção aos percentuais de dados válidos pode ter nos indicadores corporativos da indústria.

Palavras-chave: *data quality*. *python*. *data warehouse*. *data analytics*.

A Three-Dimensional Analysis of Data Quality in Purchasing Tables within a Data Warehouse Environment

ABSTRACT

In recent years, the exponential increase in data creation worldwide has drawn the attention of large companies to the importance of proper care for this valuable resource. One of the essential criteria for obtaining accurate projections and analyses of corporate data is the quality of this data, as it is crucial to ensure the integrity of the data used in any study. Although it is a relatively new topic in the market, it is crucial to consider ways to implement data quality in both small and large companies. This study aims to develop Python code to analyze three of the six recognized dimensions of data quality: completeness, compliance, and accuracy. This analysis is carried out on tables from a Data Warehouse of a Brazilian company in the retail sector. These analyses are fundamental for the proper scanning and interpretation of the databases in question, allowing for the creation of more reliable Key Performance Indicators (KPIs). The work seeks to pave the way for the improvement of data quality-related practices, demonstrate the effectiveness of applying data quality, and highlight the impact that attention to valid data percentages can have on industry corporate indicators.

Keywords: *data quality*, *python*, *data warehouse*, *data analytics*.

LISTA DE FIGURAS

LISTA DE TABELAS

Tabela 2 - Estrutura da tabela de análise.

SUMÁRIO

1 INTRODUÇÃO

A informação tornou-se um dos ativos mais cruciais para as empresas. Sem informações de alta qualidade, fornecer um serviço preciso ao cliente, tomar decisões informadas e aproveitar os benefícios das novas tecnologias se torna desafiador. O mundo contemporâneo exige informações globais e enfatiza cada vez mais a colaboração e a troca de dados tanto dentro de diferentes setores de uma empresa quanto entre diversas organizações. O impacto da correta utilização de dados confiáveis nas operações e decisões de negócios é inegável e possui uma vasta quantidade de possíveis problemas, conforme é exposto por Thomas C. Redman explica em "*Data Driven: Profiting from Your Most Important Business Asset*". Assim, pode-se entender que a vantagem competitiva de uma empresa reside no conhecimento que ela pode oferecer aos seus clientes.

Conforme Jack E. Olson explica em "*Data Quality: The Accuracy Dimension*", valorizar a aquisição e a manutenção de dados corporativos de alta qualidade possui imensa importância para as organizações, uma vez que o custo da má qualidade dos dados pode ser significativo. Dados de baixa qualidade podem resultar em falhas nos processos, custos organizacionais e até mesmo em perda parcial ou total de clientes.

Embora a busca pela melhoria dos processos sempre tenha sido uma constante, o impulso por produtividade e qualidade está levando os intervenientes da produção a alcançarem a excelência em suas saídas. Através de metodologias e técnicas, programas de qualidade e produtividade são implementados para atingir objetivos estabelecidos, conforme é discutido em "*Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*"de Danette McGilvray.

Este estudo aborda um problema de qualidade identificado dentro de uma empresa brasileira do setor varejista. A área de análise de dados detectou dificuldades na interpretação dos valores exibidos nos Indicadores-Chave de Desempenho (*KPIs*) que rastreiam os pedidos de produtos realizados por meio de uma plataforma *online*. A fonte desses painéis é o *data warehouse* corporativo que, neste caso, é a ferramenta *Snowflake*. Após analisar essas bases de dados, foram identificados problemas de qualidade dos dados nas informações que chegavam neste ambiente.

O desenvolvimento deste estudo envolveu a criação de um código *python* para leitura de dados, aplicação de métodos de qualidade dos dados e apresentação dos resultados. O motor de qualidade utilizado empregou a ferramenta *Great Expectations*, especializada em qualidade dos dados. Três dimensões da qualidade dos dados foram

examinadas: completude, conformidade e precisão.

Este trabalho demonstrará o desenvolvimento do motor de qualidade baseado em *python* e sua aplicabilidade a um cenário real — e comum — envolvendo tabelas de um *data warehouse*, fornecendo medição e análise claras de três pilares de qualidade de dados amplamente reconhecidos no mercado. Através deste estudo, a importância de manter dados corporativos de alta qualidade para a tomada de decisões estratégicas fica evidente, assim como o aumento potencial da confiabilidade dos *KPIs* corporativos quando a confiabilidade dos dados é assegurada.

Neste artigo, inicia-se com a exposição da fundamentação teórica e as tecnologias utilizadas no desenvolvimento deste projeto. A seção 2.1 expõe as diferentes vertentes do conceito de qualidade, começando pelo conceito de qualidade em si, até qualidade de dados e qualidade da informação, finalizando com o conceito de *data warehouse* o funcionamento da qualidade de dados dentro deste ambiente. A seção 2.2 apresenta o *Data Bricks* como ferramenta de *IDE*, seguido da apresentação da biblioteca *Great Expectations* com duas principais funcionalidades e aplicabilidade em qualidade de dados. Após, o capítulo 3 apresenta o estudo de caso escolhido para aplicação deste projeto de qualidade de dados. A seção 3.1 apresenta a metodologia utilizada e os pilares de qualidade de dados que foram escolhidos para serem explorados neste projeto, seguido das seções 3.2, que apresenta o domínio de dados explorado, 3.3, com o método de captura e armazenamento dos dados e, por fim, a seção 3.3, com os desafios deste projeto. O capítulo 4 apresenta a aplicação da metodologia, na seção 4.1, e as métricas de qualidade de dados adotadas na seção 4.2. Em seguida, o capítulo 5 apresenta o algoritmo proposto, e o capítulo 6 os resultados obtidos. No capítulo 7, são apresentadas as conclusões e possíveis melhorias futuras para este projeto.

2 FUNDAMENTAÇÃO TEÓRICA E TECNOLOGIAS UTILIZADAS

Este capítulo visa apresentar os principais conceitos relacionados ao contexto em que o trabalho está inserido, bem como os principais conceitos relacionados às tecnologias utilizadas. Primeiramente, na Seção 2.1, descreve-se o conceito de qualidade em geral. Em seguida, discorre-se sobre o conceito de qualidade de dados na seção 2.1.1, aplicando o conceito de qualidade ao contexto de dados, além da apresentação dos seis pilares de qualidade de dados conhecidos na literatura. Após, nas seções 2.1.2 e 2.1.3 são apresentados os conceitos de qualidade de dados em bancos de dados e o conceito de qualidade da informação. A seção seguinte, 2.1.4, apresenta o que é um *data warehouse* e como é a qualidade de dados nestes ambientes.

2.1 Conceito de Qualidade

Quando se aborda a ideia de qualidade de um produto, frequentemente se refere à sua exatidão, adequação ao uso e conformidade com um conjunto predefinido de expectativas. Também está associada à ausência de erros ou falhas no produto. A qualidade abarca os conceitos de completude e correção. De acordo com o livro "Qualidade Total: Padronização de Empresas", por Vicente Falconi, existem três conceitos fundamentais que clarificam o entendimento sobre qualidade:

- ***Requisitos de qualidade:*** São critérios ou condições cujo cumprimento qualifica um objeto em relação à sua conformidade com especificações e ao seu uso no ambiente planejado.
- ***Qualidade:*** Refere-se à conformidade com os requisitos especificados.
- ***Garantia de qualidade:*** Refere-se a atividades planejadas e sistemáticas projetadas para assegurar a confiança de que um objeto atende aos requisitos de qualidade.

2.1.1 Qualidade de Dados

A qualidade de dados é um conceito complexo que assume significados variados para diferentes indivíduos. Diversas definições foram propostas para expressar esse conceito, refletindo a natureza subjetiva da avaliação de qualidade de dados. Conforme "*Data Quality Assessment*"de Arkady Maydanchik, essa avaliação pode variar conforme a pers-

pectiva do observador, o contexto e os objetivos da análise. Logo, a qualidade muitas vezes é descrita, uma vez que nem sempre pode ser quantificada.

A avaliação das práticas de gestão de dados em uma organização requer diversas perguntas. Por exemplo, se a empresa enfrenta problemas, custos adicionais ou perdas financeiras devido à baixa qualidade dos dados. Além disso, o grau de dependência de processos automatizados de tomada de decisão e a atenção dada pela alta administração ao tratamento de dados são indicativos do comprometimento com a gestão de dados.

Para que a administração de dados seja eficaz, padrões e políticas sobre dados, sua definição e uso devem ser adotados. Esses padrões devem ser rigorosos, abrangentes e flexíveis para permitir a reutilização, estabilidade e comunicação eficaz dos dados. O uso de ferramentas como dicionários de dados e repositórios facilita a gestão dos dados.

Existem seis dimensões de qualidade de dados amplamente reconhecidas no mercado. Estas, são discutidas no artigo *"Information Quality: The Potential of Data and Analytics to Generate Knowledge"*, por Larry P. English, e servem de base fundamental para o desenvolvimento deste trabalho. São elas:

- ***Completude:*** A presença de dados completos é essencial para relatórios e indicadores precisos.
- ***Conformidade:*** Os dados devem aderir a formatos padrões e legíveis.
- ***Consistência:*** Evitar informações contraditórias entre os registros.
- ***Precisão:*** Os valores dos dados devem ser suficientemente coerentes para serem usados.
- ***Duplicação:*** Identificação de informações repetidas na mesma tabela.
- ***Integridade:*** Certificar-se de que todas as informações relevantes estão presentes e utilizáveis em um registro.

2.1.2 Qualidade de dados em Bancos de Dados

A qualidade dos dados é um requisito crítico para a interação com clientes e para decisões baseadas em soluções como *Data Warehouse*, *Data Mart* e *Data Mining*. Estes conceitos são amplamente discutidos em "Data Quality: The Accuracy Dimension"de Jack E. Olson. Neste, é possível verificar que a inexistência de dados duplicados, a visão única do cliente e a segmentação de clientes estão intrinsecamente ligados à qualidade dos dados. A avaliação pode ser quantitativa, com indicadores objetivos, ou qualitativa,

dependendo do ponto de vista do observador. Ferramentas automatizadas frequentemente auxiliam na avaliação quantitativa.

2.1.3 Qualidade da Informação

A qualidade da informação pressupõe a qualidade dos dados, do sistema de informação e do ambiente computacional. Um sistema de informação de qualidade cumpre seus objetivos, é gerenciável, mantível e compreensível por pessoas não envolvidas no projeto original. À medida que a infraestrutura de informações das empresas amadurece, aumenta a necessidade de qualidade das informações e de sistemas eficazes de suporte à decisão. Neste cenário, informações incorretas resultantes de dados não satisfatórios do ponto de vista de qualidade acabam, frequentemente, resultando em erros atualmente conhecidos como erros típicos de qualidade. Tais erros comuns são amplamente discutidos em "*Data Quality: Concepts, Methodologies and Techniques*"de Carlo Batini, et al.

2.1.4 Data Warehouse: Conceito e Qualidade de Dados

Um *data warehouse* é um tipo de sistema de gerenciamento de dados projetado para ativar e fornecer suporte às atividades de *business intelligence* (BI), especialmente a análise avançada e controle de *Key Performance Indicators* (KPIs) - indicadores-chave de desempenho, utilizados pelas organizações para medir o desempenho, o progresso e o sucesso em relação a metas e objetivos específicos. Segundo Barb Wixom em *"The current state of business intelligence in academia: The arrival of big data. Communications of the Association for Information Systems"*, a evolução do *business intelligence* está intimamente ligada aos *data warehouses*, decorrência do crescimento exponencial dos impactos do de grandes volumes de dados.

Os *data warehouses* destinam-se exclusivamente a realizar consultas e análises avançadas e geralmente contêm grandes quantidades de dados históricos. Os dados em um *data warehouse* geralmente são derivados de uma ampla variedade de fontes, como arquivos de *log* de aplicativos e aplicativos de transações. Por isso, é imprescindível realizar uma modelagem dimensional que leve em conta essa ampla variedade de fontes, como pode ser visto em *"The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Wiley."*, de Kimball, R., & Ross, M.

Um *data warehouse* centraliza e consolida grandes quantidades de dados de várias fontes. Seus recursos analíticos permitem que as organizações obtenham informações de negócios úteis de seus dados para melhorar a tomada de decisões. Com o tempo, cria-se um registro histórico que pode ser inestimável para cientistas de dados e analistas de negócios. Devido a esses recursos, um *data warehouse* pode ser considerado como a "única fonte confiável" de uma organização. O ciclo de vida da informação, considerando esses registros históricos que necessitam ser armazenados, é discutido em *"The Data Warehouse Lifecycle Toolkit. Wiley."*, de Ralph Kimball, & Margy Ross.

Um *data warehouse* típico geralmente possui alguns elementos típicos. Estes são discutidos em *"Using the Data Warehouse. Computing McGraw-Hill."*, de Inmon, W. H., & Hackathorn, R. D.. São eles:

- Um banco de dados relacional para armazenar e gerenciar dados.
- Uma solução de extração, carregamento e transformação (ELT) para preparar os dados para análise.
- Análise estatística, relatórios e recursos de mineração de dados.
- Ferramentas de análise de clientes para visualizar e apresentar dados aos usuários de negócios.
- Outras aplicações analíticas mais sofisticadas que geram informações acionáveis aplicando ciência de dados e algoritmos de inteligência artificial (IA) ou gráficos e recursos espaciais que permitem mais tipos de análise de dados em escala.

Assim, à medida que o uso do *data darehouse* se expande para informar as tomadas de decisão, é crucial adotar uma estratégia que garanta a qualidade dos dados dentro desse contexto. Dado que a qualidade dos dados tem um impacto direto nos resultados das análises, é pertinente considerar o grau de qualidade das informações analíticas durante todo o processo decisório.

Entretanto, mensurar a qualidade dos dados no ambiente do *data warehouse* não é uma tarefa trivial, principalmente quando a qualidade dos processos está entrelaçada com os resultados obtidos. O ambiente de um *data warehouse* desempenha um papel fundamental ao disseminar o conhecimento do negócio, culminando em uma inteligência competitiva. Nesse cenário, de acordo com Larry P. English em "*Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*", a introdução de medidas que garantam a qualidade dos dados torna-se indispensável, transformando a qualidade de dados em uma pedra angular do *data warehouse*. A

necessidade de consistência na qualidade dos dados cresce proporcionalmente ao aumento da complexidade e do volume das fontes de informações.

Uma abordagem abrangente em relação à qualidade dos dados exige a evolução constante da qualidade dos valores dos dados, alcançada por meio de Verificação, Validação e Certificação (VVC), juntamente com a evolução dos processos que geram e modificam os dados, conforme, W. H. Inmon explica em "*Building the Data Warehouse*". Sendo o foco aumentar a qualidade dos dados produzidos, é possível utilizar ferramentas e tecnologias como aliados nesse processo. Uma alta gama de ferramentas de qualidade de dados pode ser vista em "*The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*"de Ralph Kimball, bem como de que forma elas se encaixam no processo da manutenção de qualidade de dados corporativa.

No contexto da consistência dos dados em um *data warehouse*, determinados aspectos ganham destaque: integridade, precisão e completude. A integridade garante a segurança dos dados armazenados na fonte. A precisão reflete o quão fielmente os dados da fonte representam a realidade. A completude avalia o grau em que todos os dados necessários para atender às demandas de negócio estão disponíveis na fonte. Ao analisar como uma organização adota práticas de gestão de dados, uma série de indagações emerge. Uma delas considera se a empresa enfrenta problemas, custos adicionais ou perdas financeiras que surgem devido à baixa qualidade de seus dados, expondo-a a riscos potenciais.

Frequentemente, a qualidade dos dados de uma organização se torna uma vantagem competitiva. Nesses casos, é possível rapidamente identificar oportunidades de negócio ou *marketing* por meio da transformação e análise de dados comerciais atuais e históricos.

O crescente uso do *data warehouse* para sustentar processos decisórios tem gerado um aumento na preocupação em relação à qualidade dos dados. A probabilidade de aproveitar os recursos de informação é ampliada quando há um entendimento prévio da confiabilidade desses recursos.

2.2 Ferramentas Utilizadas

Neste capítulo, são exploradas as tecnologias que serviram como ferramentas fundamentais para a realização deste trabalho. Na primeira seção, é apresentado o *DataBricks*, utilizado principalmente como uma plataforma de desenvolvimento integrada

(IDE) para executação do código *python*. Na segunda seção, é destacada a biblioteca *Great Expectations*, uma ferramenta dedicada à exploração de qualidade de dados.

2.2.1 *Data Bricks* como ferramenta de IDE

Para o desenvolvimento, foi utilizada a plataforma *Databricks*. Esta é uma plataforma de análise e processamento de dados baseada em nuvem, projetada para simplificar a implementação e o gerenciamento de fluxos de trabalho de *big data* e análise de dados. Ela fornece uma série de ferramentas e serviços que facilitam a colaboração entre equipes, o desenvolvimento de código, a execução de análises avançadas e a criação de *pipelines* de dados escaláveis.

Uma das características principais do *Databricks* é sua funcionalidade de "IDE de Programação"(Ambiente de Desenvolvimento Integrado), que oferece um espaço centralizado e amigável para desenvolver, testar e executar código em várias linguagens, como *Python*, *Scala*, SQL e R. Por ser um ambiente interativo para desenvolver, testar e executar código para análise de dados e por já possuir integração nativa com a biblioteca *Great Expectations*, foi escolhida como plataforma de desenvolvimento deste projeto.

2.2.2 Avaliação da Qualidade de Dados com a biblioteca *Great Expectations*

A biblioteca *Great Expectations* foi a escolha central para avaliar a qualidade dos dados devido à sua especialização em qualidade de dados e a capacidade de lidar com a complexidade das métricas de conformidade, completude e precisão, dimensões escolhidas para serem estudadas neste projeto.

Esta é uma ferramenta de código aberto desenvolvida para facilitar a avaliação e o monitoramento da qualidade de dados em projetos de análise de dados e integração de dados. A biblioteca oferece uma abordagem sistemática e programática para definir, testar e documentar as expectativas sobre os dados, permitindo que os usuários identifiquem rapidamente problemas de qualidade e tomem medidas corretivas.

A *Great Expectations* permite que os usuários definam expectativas sobre seus dados por meio de configurações simples e declarativas. Essas expectativas representam critérios específicos que os dados devem atender, como valores não nulos, formatos específicos, intervalos de valores aceitáveis, entre outros. Permite, também, a execução

automatizada de verificações de qualidade de dados em relação às expectativas definidas. Isso pode ser feito em diferentes pontos do fluxo de dados, como durante a ingestão, transformação ou antes da análise. As verificações são realizadas por meio de comandos simples, permitindo que os usuários monitorem continuamente a qualidade dos dados. Ela também é flexível e pode ser usada para verificar a qualidade de dados em várias fontes, como bancos de dados relacionais, *data lakes*, serviços *web* e planilhas, entre outros.

3 ANÁLISE DA TABELA *ORDER DETAILS TABLE*

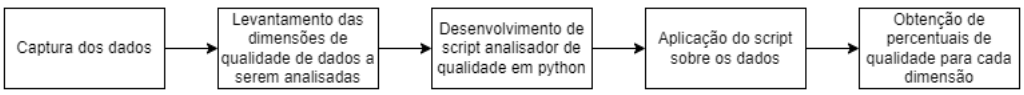
Neste capítulo, é apresentado o conjunto de dados empregado para a realização das análises de qualidade e seu objetivo, bem como a metodologia utilizada ao decorrer do projeto. Por razões de confidencialidade, a empresa optou por não divulgar seu nome.

3.1 Metodologia

A figura 3.1 ilustra a abordagem metodológica adotada e consistiu na metodologia de estudo de caso. Nesse sentido, uma análise minuciosa e aprofundada foi conduzida em um caso específico, identificado em uma empresa brasileira do setor de varejista. Essa análise permitiu alcançar uma compreensão abrangente do fenômeno relacionado à qualidade de dados. Consequentemente, o escopo deste trabalho buscou desenvolver um projeto que visasse avaliar os indicadores de qualidade de dados nas dimensões de completude, conformidade e precisão dentro do âmbito do domínio de dados selecionado. O resultado final desejado consiste em percentuais de completude, de conformidade e de precisão do domínio de dados. Dessa forma, os passos seguidos para obtenção dos percentuais de qualidade de dados para as dimensões selecionadas sobre os dados escolhidos foram:

1. Captura dos dados
2. Levantamento das dimensões de qualidade de dados a serem analisadas
3. Desenvolvimento de *script* analisador de qualidade em *python*
4. Aplicação do *script* sobre os dados
5. Obtenção de percentuais de qualidade para cada dimensão

Figura 3.1: Fluxograma de trabalho.



3.2 Domínio de Dados

A empresa cujos dados foram utilizados para o desenvolvimento deste trabalho é uma empresa do comércio varejista. Por questões de confidencialidade, a empresa preferiu manter-se anônima. Esta possui uma área dedicada a análise de dados, que levantou a necessidade de analisar a qualidade dos dados de um domínio específico de dados após relatar problemas em relatórios do *PowerBI*. O domínio de dados refere-se ao conjunto de informações relacionadas a pedidos *online*. Possui informações como, por exemplo, identificador do produto comprado, quantidade de produtos e local em que o pedido foi realizado.

Todo esse conjunto de dados é armazenado em uma tabela dentro do *data warehouse Snowflake*, a qual existe, na mesma estrutura, replicada para onze países. São eles: África do Sul, Argentina, México, Bolívia, Paraguai, Colombia, Panamá, República Dominicana, Honduras, El Salvador e Peru. Cada país possui seu próprio *database* dentro do *data warehouse* e, portanto, a mesma estrutura de tabela onde as informações dos pedidos são armazenadas.

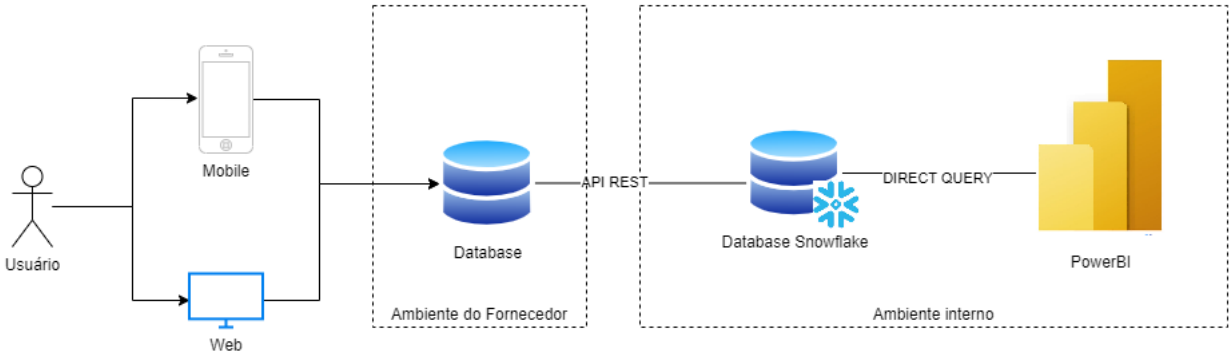
A tabela possui 169 colunas. Para o escopo deste estudo, foram selecionadas 18 colunas para análise de qualidade. Podemos identificar a estrutura da Tabela 2.

3.3 Captura e Armazenamento de Dados

A origem dos dados analisados é uma plataforma *web* de pedidos, onde os clientes realizam solicitações através de uma *interface* específica para este contexto. Os pedidos são realizados via *interface web* ou aplicativo *mobile*. Essas informações são capturadas e armazenadas pelo fornecedor contratado para esta solução e, portanto, não está presente neste trabalho detalhes sobre seu funcionamento interno.

Após, através de *API REST*, as informações desses pedidos chegam até o *data warehouse Snowflake* dentro do ambiente corporativo. O *data warehouse* armazena os dados em sua tabela estruturada especificamente para este escopo, replicada em onze diferentes bases de dados relativos a cada país, e o *PowerBI* corporativo conecta nestas bases via *direct query*. Sobre o *PowerBI*, os analistas de dados produzem seus relatórios. Na figura 3.2, é possível identificar a arquitetura do processo de captura, armazenamento e análise dos dados.

Figura 3.2: Arquitetura de captura e armazenamento de dados.



Este fluxograma existe desta maneira para cada um dos onze países, que são o total de clientes da empresa aqui utilizada como material de estudo. As tabelas são constantemente alimentadas com novos pedidos *online*. Para a análise deste trabalho, foram capturados dados de quatro dias diferentes da semana, sendo eles: dia 14 de agosto, segunda-feira, dia 16 de agosto, quarta-feira, dia 18 de agosto, sexta-feira e dia 19 de agosto, sábado, todos em 2023. Os dias da semana foram escolhidos randomicamente. Foi decidido analisar dados de diferentes dias da semana para verificar se havia alguma inconsistência a depender do dia da semana.

3.4 Desafios

O objetivo central deste estudo é calcular os índices de qualidade de dados, permitindo avaliar com maior confiança o grau de confiabilidade dos dados corporativos finais. Ao analisar completude, conformidade e precisão dos dados, a empresa pode embasar decisões mais informadas e considerar ajustes na arquitetura da solução de coleta, armazenamento e análise de dados, conforme é explorado em "*Data Quality: A Survival Guide*"de Tom Redman, que possui casos reais semelhantes ao escolhido neste projeto.

Cabe ressaltar que este trabalho não tem o propósito de corrigir falhas na qualidade, uma vez que tal tarefa exige revisão das integrações com os fornecedores da solução de captura de dados e consideração das particularidades de cada país. No entanto, os índices obtidos contribuem para aprimorar a compreensão das informações derivadas desses dados. Assim, a meta é alcançar índices de completude, conformidade e precisão para cada país em cada dia da semana.

Considerando-se o contexto das informações das colunas, foram selecionadas colunas específicas para a aplicação da análise de cada uma das dimensões de qualidade,

Coluna	Descrição	Tipo	Exemplo
BUSINESS SPK	Nome do domínio de dados no país	VARCHAR(22)	mx modelorama ondemand
DELIVERY ADDRESS SPK	ID do endereço do pedido	VARCHAR(16777216)	b943e7c61bee4e758085e87679d3d8f3
DELIVERY ADDRESS STATE	Estado do pedido	VARCHAR(16777216)	Jalisco
DELIVERY TYPE	Modo de entrega	VARCHAR(16777216)	trem
ITEM NAME	Nome do produto	VARCHAR(16777216)	pepsi black
ITEM SKU	ID do item no cliente comercial	VARCHAR(255)	751.00000
ORDER AK	ID do pedido na base interna	VARCHAR(16777216)	1162401678404
ORDER DATE	Data do pedido	Date	2022-09-12
ORDER NK	ID do pedido na base fonte	VARCHAR(16777216)	1162401678404-01
ORDER SPK	ID do setor do pedido + ID pedido	VARCHAR(16777216)	O1524;1316593366959-01
ORDER STATUS TYPE	Status do pedido	VARCHAR(16777216)	Valid
ORDER TYPE	Tipo de compra	VARCHAR(10)	Normal
PRODUCT NK	ID do produto	VARCHAR(255)	244
PRODUCT SPK	ID do setor do produto + ID produto	VARCHAR(255)	O1524;553
SOURCE	Nome da fonte criadora dos dados	VARCHAR(22)	mx ondemand
UNIT PRICE OFFERED TAX EXCL LOCAL	Preço unitário do produto	FLOAT	34,99
UNIT PRICE LIST TAX EXCL LOCAL	Valor do imposto local	FLOAT	15,36
UNIT QUANTITY	Número de unidades do produto	NUMBER(38,0)	3

:

Tabela 2

ou seja, cada métrica de qualidade teve seu conjunto específico de colunas para análise. Dessa forma, cada país teve, em cada um dos quatro dias da semana, e para cada dimensão analisada, percentuais de qualidade para cada uma das colunas, bem como um percentual geral de quantidade total de registros corretos do dia (considerando todos os dados de todas as colunas analisadas). Ao final de cada semana, registrou-se, também, um percentual geral final de qualidade de dados para cada dimensão, correspondente à quantidade total de registros corretos deste país nos quatro dias de análise.

4 APLICAÇÃO DA METODOLOGIA

Este capítulo tem como objetivo apresentar a aplicação da metodologia selecionada, alinhada com as dimensões de qualidade que estão sendo investigadas. Além disso, serão discutidas as tecnologias empregadas para obter os resultados desejados.

4.1 Visão Geral da Metodologia

Após interações com a equipe de analistas de dados e com uma análise visual das bases de dados, foram identificadas três hipóteses relativas a problemas de qualidade. A primeira envolvia a ausência de dados, ou seja, colunas com registros não preenchidos em sua totalidade. A segunda se referia a dados que não estavam em conformidade com o padrão de dados esperado, incluindo campos destinados a quantidades de produtos (do tipo numérico) preenchidos com caracteres alfabéticos, por exemplo. A terceira hipótese abordava informações inconsistentes, como a presença de informações de estados que, na verdade, não existiam no contexto de um país. Essas observações levaram à escolha de três dimensões de qualidade de dados para serem exploradas: completude, conformidade e precisão, que serão detalhadas e desenvolvidas na seção a seguir.

Para o desenvolvimento do trabalho, foi realizada a criação de um código em *python* que extraiu os dados das tabelas do *Snowflake* nos dias determinados e procedeu à análise das colunas conforme as três dimensões selecionadas. Para essa tarefa, a IDE escolhida foi o *DataBricks*, e a principal biblioteca de qualidade de dados utilizada foi o *Great Expectations*, que serão apresentados nas seções seguintes.

4.2 Métricas de qualidade de dados adotadas

Como base teórica para a análise de qualidade deste projeto, foram adotadas três das seis dimensões conhecidas de qualidade de dados. Estas, foram interpretadas para o contexto dos dados em questão e medidas de acordo com as corretas *features* da biblioteca *Great Expectations*.

4.2.1 Métrica 1 - Completude

A completude dos dados envolve a presença de registros com valores em branco ou "*null*". Esta verificação foi aplicada a todas as colunas das tabelas do projeto, pois, sendo todas informações chave para os relatórios e *KPIs* no *PowerBI*, necessitavam que seus registros viessem preenchidos. Para esta verificação, foi empregada a *feature* "*expect column values to not be null*"da biblioteca *Great Expectations*, que verifica o conteúdo das colunas lidas.

4.2.2 Métrica 2 - Conformidade

A conformidade dos dados se refere à presença de campos obrigatórios com formatos incorretos. Por exemplo, um campo de ID que deve seguir o formato "12.345"precisa receber registros de dois números, seguidos pelo caractere “.”, seguido de mais três números.

Para esta análise, foi utilizada a *feature* da biblioteca *Great Expectations* "*expect column values to match regex*". Esta *feature* compara o conteúdo dos registros de análise com expressões regulares previamente definidas.

Esta métrica foi aplicada a quatorze colunas, cada uma com a sua própria expressão regular definida. As colunas medidas em conformidade e suas expressões regulares podem ser observadas na tabela 4.1.

4.2.3 Métrica 3 - Precisão

A dimensão de precisão foi aplicada especificamente para medir a qualidade da informação do estado do país em que o pedido foi realizado, presente na coluna "*DELIVERY ADDRESS STATE*". Os registros com estados inexistentes não são considerados corretos.

Foi utilizada, para a avaliação de precisão dos dados, a *feature* "*expect values to be in set*"da biblioteca *Great Expectations*. Sua funcionalidade é baseada na comparação dos registros lidos com valores pré definidos, armazenados em variáveis do tipo lista. Dessa forma, registros escritos errados ou fora do padrão de conformidade acabam por serem considerados inválidos também.

Coluna	REGEX	Exemplo
BUSINESS SPK	"Ĉ(?= . * [A-Za-z])[A-Za-z0-9]+ $ "	mx modelorama ondemand
DELIVERY ADDRESS SPK	"Ĵ[0-9a-fA-F]8(-[0-9a-fA-F]4)3-[0-9a-fA-F]12 $ "	b943e7c61bee4e758085e87679d3d8f3
DELIVERY ADDRESS STATE	"\w+"	Jalisco
ITEM NAME	"(\d+) \s*(?:packlPack)? \s+([A-Za-z \s]+)"	pepsi black
ITEM SKU	"\d3 \. \d5"	751.00000
ORDER AK	"\d+"	1162401678404
ORDER DATE	"\d4- \d2- \d2"	2022-09-12
ORDER SPK	"O \d4; \d+- \d2"	O1524;1316593366959-01
ORDER TYPE	"Ĵ[A-Za-z]+ $ "	Normal
PRODUCT NK	"\d+"	244
PRODUCT SPK	"O \d4; \d+"	O1524;553
UNIT PRICE OFFERED TAX EXCL LOCAL	"\d+"	34,99
UNIT PRICE LIST TAX EXCL LOCAL	"\d+"	15,36
UNIT QUANTITY	"\d+"	3

Tabela 4.1: Tabela contendo as colunas analisadas, suas expressões regulares e exemplos.

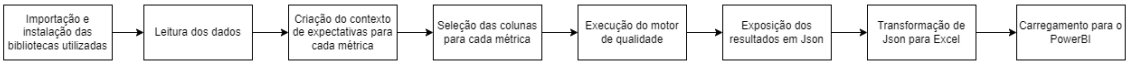
5 DESENVOLVIMENTO

Este capítulo tem como propósito detalhar o desenvolvimento do projeto, incluindo a adaptação da abordagem metodológica ao contexto do domínio de dados, fazendo uso das tecnologias destacadas.

5.1 Visão Geral

O algoritmo de alto nível aplicado é composto pelas seguintes etapas: importação e instalação das bibliotecas utilizadas, leitura dos dados, criação do contexto de expectativas para cada métrica, seleção das colunas para cada métrica, execução o motor de qualidade, exposição dos resultados em *json*, transformação de *json* para Excel, carregamento para o PowerBI. Este fluxo é ilustrado na Figura 5.1.

Figura 5.1: Algoritmo alto nível.



Para exibir os resultados, foram desenvolvidos painéis interativos no *Power BI*. Serão mostrados os percentuais de qualidade de dados para cada dimensão, em relação a cada dia da semana, utilizando a África do Sul como país de exemplo para a dimensão de completude, o país Bolívia como exemplo dos resultados para a dimensão de conformidade e o país Panamá como exemplo de análise para a dimensão de precisão.

5.2 Algoritmo Proposto

Nesta seção, serão exibidos os segmentos do código *python* elaborado para este projeto. Será apresentado o código correspondente a cada uma das etapas do algoritmo, destacando as tecnologias empregadas em cada caso.

1. Importação e instalação das bibliotecas utilizadas:
   Instalação da biblioteca *Great Expectation* e importação do módulo "*great expectations*"e da classe "*checkpoint*".

Figura 5.2: Importação e instalação das bibliotecas utilizadas.

```
pip install great_expectations
import great_expectations as gx
from great_expectations.checkpoint import Checkpoint
```

Importação da biblioteca *Spark* para manipulação de *dataframes*:

Figura 5.3: Importação da biblioteca *Spark*.

```
from pyspark.sql import SparkSession
```

2. Leitura dos dados:

   Carregamento dos dados do *Snowflake* via consulta SQL.

Figura 5.4: Carregamento dos dados via consulta SQL.

```
sf = SnowflakeIntegrator('<database>', '<access_role>')
df = sf.loadQuery('SELECT * FROM <database>.<country_schema>.<order_table>')
```

Criação de um *dataframe*:

Figura 5.5: Criação de um *dataframe*.

```
dataframe_datasource = context.sources.add_or_update_spark(
    name="my_spark_in_memory_datasource",
)
```

Carregamento dos dados para o *dataframe*.

Figura 5.6: Carregamento dos dados para o *dataframe*.

```
dataframe_asset = dataframe_datasource.add_dataframe_asset(
    name="yellow_tripdata",
    dataframe=df,
)
```

3. Criação do contexto de expectativas para cada métrica:

   Construção de uma solicitação de lote (*batch request*):

Figura 5.7: Criação do *batch request*.

```
batch_request = dataframe_asset.build_batch_request()
```

Configuração do contexto como uma raíz de diretório, local onde as configurações, expectativas e outros artefatos serão armazenados:

Figura 5.8: Criação do *data context*.

```
context_root_dir = "/dbfs/great_expectations/"
context = gx.get_context(context_root_dir=context_root_dir)
```

Criação da suíte de expectativas para completude:

Figura 5.9: Suíte de expectativa: completude.

```
expectation_suite_name = "expect_column_values_to_not_be_null"
context.add_or_update_expectation_suite(expectation_suite_name=expectation_suite_name)
validator = context.get_validator(
    batch_request=batch_request,
    expectation_suite_name=expectation_suite_name,
)
```

Criação da suíte de expectativas para conformidade:

Figura 5.10: Suíte de expectativa: conformidade.

```
expectation_suite_name_regex = "expect_column_values_to_match_regex"
context.add_or_update_expectation_suite(expectation_suite_name=expectation_suite_name_regex)
validator_regex = context.get_validator(
    batch_request=batch_request,
    expectation_suite_name=expectation_suite_name,
)
```

Criação da suíte de expectativas para precisão:

Figura 5.11: Suíte de expectativa: precisão.

```
expectation_suite_name_in_set = "expect_column_values_to_be_in_set"
context.add_or_update_expectation_suite(expectation_suite_name=expectation_suite_name_in_set)
validator_in_set = context.get_validator(
    batch_request=batch_request,
    expectation_suite_name=expectation_suite_name,
)
```

4. Seleção das colunas para cada métrica: Completude:

Figura 5.12: Seleção das colunas para completude.

```
# COMPLETUDE: valores não null
validator.expect_column_values_to_not_be_null(column="SOURCE")
validator.expect_column_values_to_not_be_null(column="ORDER_NK")
validator.expect_column_values_to_not_be_null(column="ORDER_AK")
validator.expect_column_values_to_not_be_null(column="ORDER_SPK")
validator.expect_column_values_to_not_be_null(column="BUSINESS_SPK")
validator.expect_column_values_to_not_be_null(column="PRODUCT_NK")
validator.expect_column_values_to_not_be_null(column="ITEM_SKU")
validator.expect_column_values_to_not_be_null(column="ITEM_NAME")
validator.expect_column_values_to_not_be_null(column="PRODUCT_SPK")
validator.expect_column_values_to_not_be_null(column="ORDER_DATE")
validator.expect_column_values_to_not_be_null(column="ORDER_TYPE")
validator.expect_column_values_to_not_be_null(column="UNIT_QUANTITY")
validator.expect_column_values_to_not_be_null(column="UNIT_PRICE_LIST_TAX_EXCL_LOCAL")
validator.expect_column_values_to_not_be_null(column="UNIT_PRICE_OFFERED_TAX_EXCL_LOCAL")
validator.expect_column_values_to_not_be_null(column="DELIVERY_ADDRESS_SPK")
validator.expect_column_values_to_not_be_null(column="ORDER_STATUS_TYPE")
validator.expect_column_values_to_not_be_null(column="DELIVERY_TYPE")
validator.expect_column_values_to_not_be_null(column="DELIVERY_ADDRESS_STATE")
```

Conformidade:

Figura 5.13: Seleção das colunas para conformidade.

```
# CONFORMIDADE: valores no formato correto
validator.expect_column_values_to_match_regex(column="ORDER_AK", regex = "\d+")
validator.expect_column_values_to_match_regex(column="ORDER_SPK", regex = "O\d{4};\d+-\d{2}")
validator.expect_column_values_to_match_regex(column="BUSINESS_SPK", regex = "^(?=.*[A-Za-z])[A-Za-z0-9]+$")
validator.expect_column_values_to_match_regex(column="PRODUCT_NK", regex = "\d+")
validator.expect_column_values_to_match_regex(column="ITEM_SKU", regex = "\d{3}\.\d{5}")
validator.expect_column_values_to_match_regex(column="ITEM_NAME", regex = "(\d+)\s*(?:pack|Pack)?\s+([A-Za-z\s]+)")
validator.expect_column_values_to_match_regex(column="PRODUCT_SPK", regex = "O\d{4};\d+")
validator.expect_column_values_to_match_regex(column="ORDER_DATE", regex = "\d{4}-\d{2}-\d{2}")
validator.expect_column_values_to_match_regex(column="ORDER_TYPE", regex = "^[A-Za-z]+$")
validator.expect_column_values_to_match_regex(column="UNIT_QUANTITY", regex = "\d+")
validator.expect_column_values_to_match_regex(column="UNIT_PRICE_LIST_TAX_EXCL_LOCAL", regex = "\d+")
validator.expect_column_values_to_match_regex(column="UNIT_PRICE_OFFERED_TAX_EXCL_LOCAL", regex = "\d+")
validator.expect_column_values_to_match_regex(column="DELIVERY_ADDRESS_SPK", regex = "^[0-9a-fA-F]{8}(-[0-9a-fA-F]{4}){3}-[0-9a-fA-F]{12}$")
validator.expect_column_values_to_match_regex(column="DELIVERY_ADDRESS_STATE", regex = "\w+")
```

Precisão:

Definição das listas de valores satisfatórios para os estados de cada país:

Figura 5.14: Lista de estados do México.

```
# México
value_set_1 = ['Chihuahua', 'chihuahua', 'Sonora', 'sonora', 'Coahuila', 'coahuila', 'Durango', 'durango',
'Oaxaca', 'oaxaca', 'Jalisco', 'jalisco',
                'Tamaulipas', 'tamaulipas', 'Chiapas', 'chiapas', 'Baja California Sur', 'baja california sur',
                'Zacatecas', 'zacatecas', 'Veracruz', 'veracruz',
                'Baja California', 'baja california', 'Nuevo León', 'nuevo león', 'Guerrero', 'guerrero', 'San
                Luis Potosí', 'san luis potosi', 'michoacán']
```

5. Execução do motor de qualidade:

   A suíte de expectativas é salva com as expectativas definidas anteriormente. É setado como "*false*" o argumento "*discard failed expectations*" para que as expectativas que falharam se mantenham na suíte para análise posterior:

Figura 5.15: Salvando a suite de expectativas no *validator*.

```
validator.save_expectation_suite(discard_failed_expectations=False)
```

É criado o *checkpoint*, responsável por criar pontos de controle em *pipelines* de verificação de dados. É o que permite o monitoramento e rastreabilidade das verificações de expectativas realizadas:

Figura 5.16: Criando o *checkpoint*.

```
my_checkpoint_name = "my_databricks_checkpoint"

checkpoint = Checkpoint(
    name=my_checkpoint_name,
    run_name_template="jr-my-run-name-template",
    data_context=context,
    batch_request=batch_request,
    expectation_suite_name=expectation_suite_name,
    action_list=[
        {
            "name": "store_validation_result",
            "action": {"class_name": "StoreValidationResultAction"},
        },
        {"name": "update_data_docs", "action": {"class_name": "UpdateDataDocsAction"}},
    ],
)
```

É adicionado o ponto de verificação criado ao contexto do *Great Expectations* e acionamos. Nesse ponto, as validações são realizadas com base nas expectativas definidas na suíte e nos dados do lote expecificado:

Figura 5.17: Criação do ponto de verificação.

```
context.add_or_update_checkpoint(checkpoint=checkpoint)
checkpoint_result = checkpoint.run()
```

6. Exposição dos resultados em *json*:

   Os resultados da execução do ponto de verificação são acessados e armazenados na variável "*run results*". Esses resultados incluem as informações sobre as expectativas que passaram, falharam ou foram ignoradas, bem como outras métricas de validação. O resultado é, então, entregue no formato *json*:

Figura 5.18: Exemplo de arquivo *json*.

```
{
  "run_id": {
    "run_name": "jr-my-run-name-template",
    "run_time": "2023-08-19T14:40:41.649421+00:00"
  },
  "run_results": {
    "ValidationResultIdentifier::expect_column_values_to_not_be_null/jr-my-run-name-template/20230819T144041.649421Z/my_spark_in_memory_datasource-yellow_tripdata": {
      "validation_result": {
        "success": false,
        "results": [
          {
            "success": true,
            "expectation_config": {
              "expectation_type": "expect_column_values_to_not_be_null",
              "kwargs": {
                "column": "SOURCE",
                "batch_id": "my_spark_in_memory_datasource-yellow_tripdata"
              },
              "meta": {}
            },
            "result": {
```


O *json* possui registros que guardam as informações: métrica de qualidade sendo testada, coluna que está sendo avaliada, número de registros (linhas da tabela) lidos para avaliação e número de registros fora do padrão setado.

Figura 5.19: Exemplo de conteúdo do arquivo *json*.


```
{
  "success": true,
  "expectation_config": {
    "expectation_type": "expect_column_values_to_not_be_null",
    "kwargs": {
      "column": "SOURCE"
      "batch_id": "my_spark_in_memory_datasource-yellow_tripdata"
    },
    "meta": {}
  },
  "result": {
    "element_count": 5618445,
    "unexpected_count": 0,
    "unexpected_percent": 0.0,
    "partial_unexpected_list": [],
    "partial_unexpected_counts": []
  }
```


7. Transformação de Json para Excel:

   Esta etapa foi realizada manualmente com o auxílio do *site* gratuito *ttps://products.aspose.app/cells/pt to-xlsx*.

   Cada dia em que o algoritmo foi executado produziu um arquivo *json* para cada país, que, consequentemente, foi transformado em um arquivo excel. Ou seja, no final de cada dia de execução do algoritmo, foram produzidos onze arquivos *json* e onze arquivos do tipo excel.

   Manualmente, para cada dia, os arquivos excel foram compilados em um só e, no final da semana, os quatro arquivos excel foram compilados para um só, guardando a marcação de data.

Dessa forma, o arquivo final excel contém as informações:

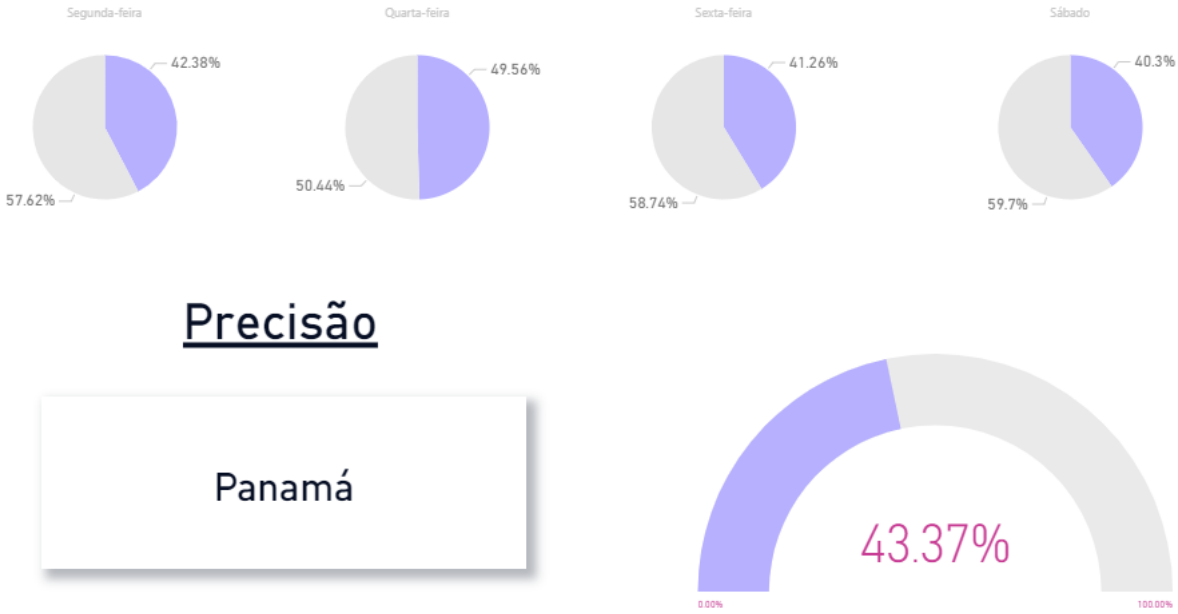
- *Country*: Nome do país, adicionado manualmente após cada execução individual do algoritmo;
- *Column*: Nome da coluna avaliada, extraída da saída do algoritmo.
- *Criteria*: Critério de qualidade sendo avaliado, extraído da saída do algoritmo.
- *Date*: Dia da semana em que a informação foi capturada, adicionada manualmente.
- *Element count*: Número de registros lidos para avaliação, extraído da saída do algoritmo.
- *Unexpected values* Número de registros com valor diferente do esperado, extraído da saída do algoritmo.

Figura 5.20: Exemplo da tabela final excel.

Country	Column	Criteria	Date	Element count	Unexpected values
Mexico	SOURCE	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	ORDER_NK	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	ORDER_AK	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	ORDER_SPK	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	BUSINESS_SPK	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	PRODUCT_NK	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	ITEM_SKU	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	ITEM_NAME	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	PRODUCT_SPK	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	ORDER_DATE	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	ORDER_TYPE	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	UNIT_QUANTITY	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	UNIT_PRICE_LIST_TAX_EXCL_I	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	UNIT_PRICE_OFFERED_TAX_E	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	DELIVERY_ADDRESS_SPK	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	ORDER_STATUS_TYPE	expect_column_values_to_not_be_null	segunda	5618445	0
Mexico	DELIVERY_TYPE	expect_column_values_to_not_be_null	segunda	5618445	5618445
Mexico	DELIVERY_ADDRESS_STATE	expect_column_values_to_not_be_null	segunda	5618445	0

8. Carregamento para o *PowerBI*: Através da conexão nativa entre o *PowerBI* e o Excel, os dados foram carregados para o *PowerBI* e expostos conforme o exemplo de *dashboards* a seguir:

Figura 5.21: Exemplo de *dashboard* no *PowerBI* - dados de precisão do Panamá.



6 RESULTADOS

Neste capítulo, apresentaremos os resultados derivados da execução do algoritmo. A fim de ilustrar esses resultados, utilizaremos os países África do Sul, como exemplo de qualidade de dados em completude, Bolívia, em conformidade, e Panamá em precisão.

As análises de resultado foram criadas sobre os atributos:

- ***Country:*** País referente dos dados.
- ***Column:*** Coluna analisada.
- ***Criteria:*** Critério de qualidade analisado.
- ***Date:*** Dia da semana em que os dados foram capturados.
- ***Element Count:*** Qualidade total de registros.
- ***Unexpected Values:*** Registros incorretos de acordo com o "Criteria"selecionado.

Os resultados foram analisados separadamente para cada métrica de qualidade abordada neste projeto. Para cada métrica de qualidade, serão demonstrados os percentuais de qualidade de cada uma das colunas analisadas em cada dia da semana, bem como um percentual geral de cada dia da semana, correspondente ao percentual total de registros válidos do dia, equivalente à média de qualidade de todas as colunas.

Ao final das análises de todos os dias de cada métrica, é, também, apresentado um percentual final de qualidade, correspondente à média dos valores diários de qualidade do conjunto de dados em questão. Ou seja, para cada métrica de qualidade, existirão quatro percentuais diários e um percentual final de qualidade.

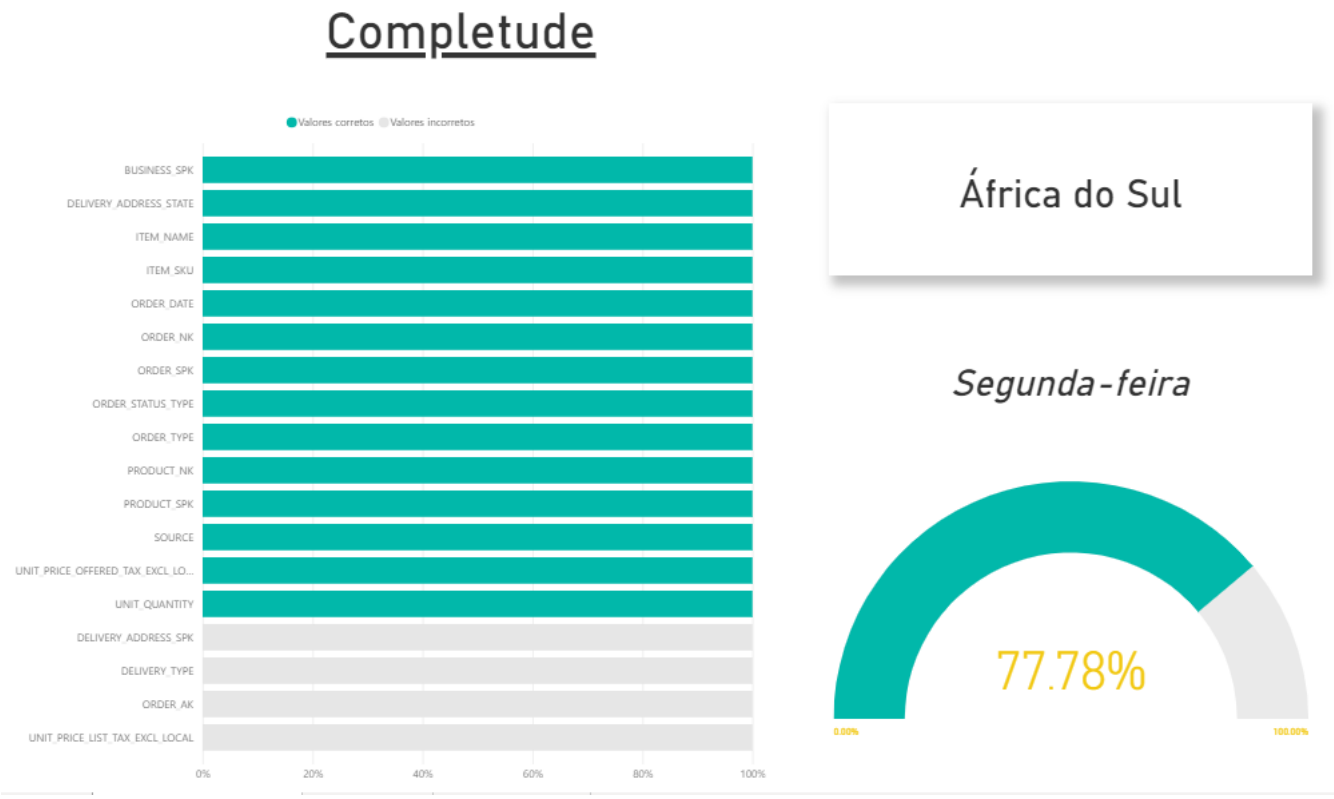
6.1 Análise de Completude

A análise de completude busca observar se os registros de cada coluna possuem dados diferentes de "*null*". A cada coluna, espera-se um percentual de registros que estão de acordo com a completude de qualidade, ou seja, que receberam dados diferentes de "*null*"a cada dia para o conjunto de dados em questão. A análise de completude foi realizada sobre todas as dezoito colunas do domínio de dados.

Cada dia da semana possui, também, um percentual total de completude, correspondente à média dos percentuais de suas colunas. Os valores de completude para o país África do Sul nos dias correspondentes a segunda-feira, quarta-feira, sexta-feira e sábado, podem ser vistos a seguir:

- Segunda-feira: 77,78 %

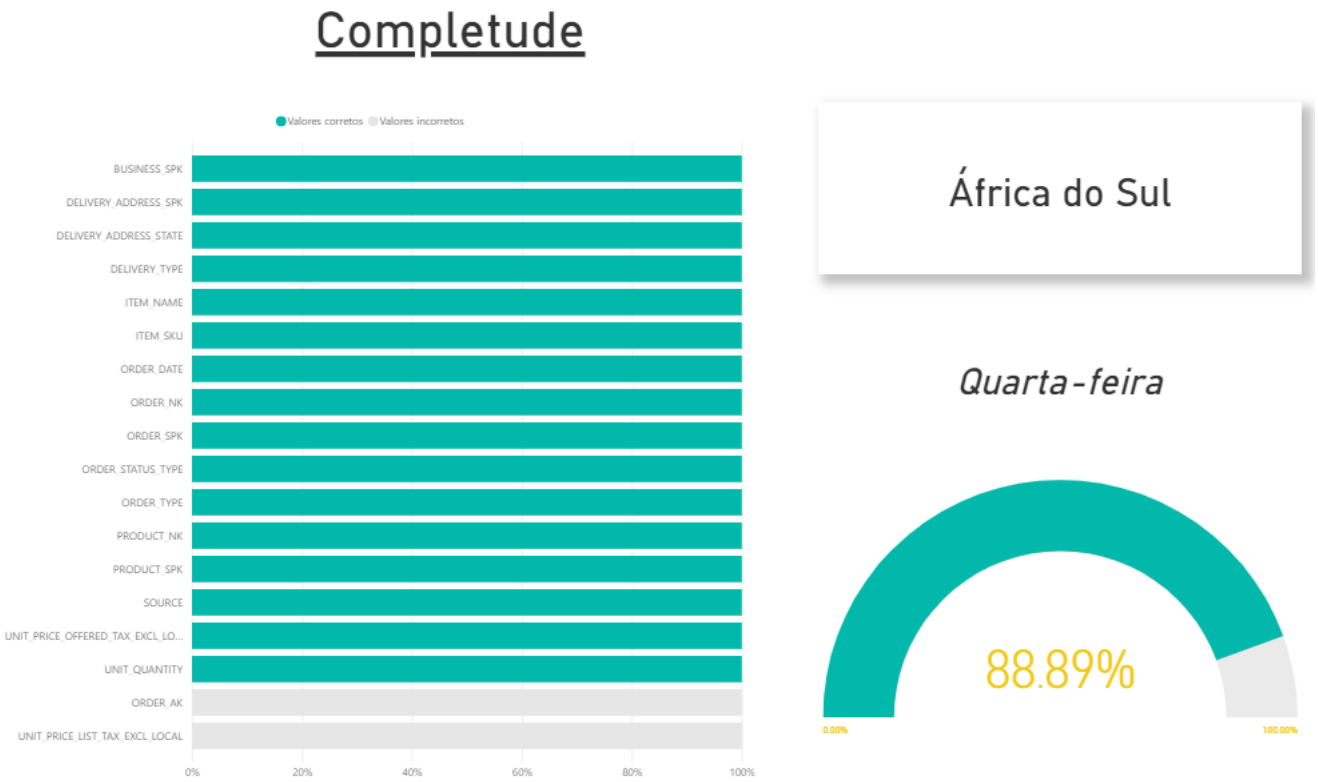
Figura 6.1: África do Sul - Completude, segunda-feira.



Na segunda-feira, a África do Sul obteve 77,78% das suas colunas de de acordo com a completude da qualidade de dados. É possível observar que 14 das 18 colunas vieram com registros diferentes de "*null*", ou seja, com resultados satisfatórios. As colunas *"DELIVERY ADDRESS SPK"*, *"DELIVERY TYPE"*, *"ORDER AK"* e *"UNIT PRICE LIST TAX EXCL LOCAL"* vieram completamente corrompidas, enquanto as outras todas estiveram totalmente coerentes.

- Quarta-feira: 88,89 %

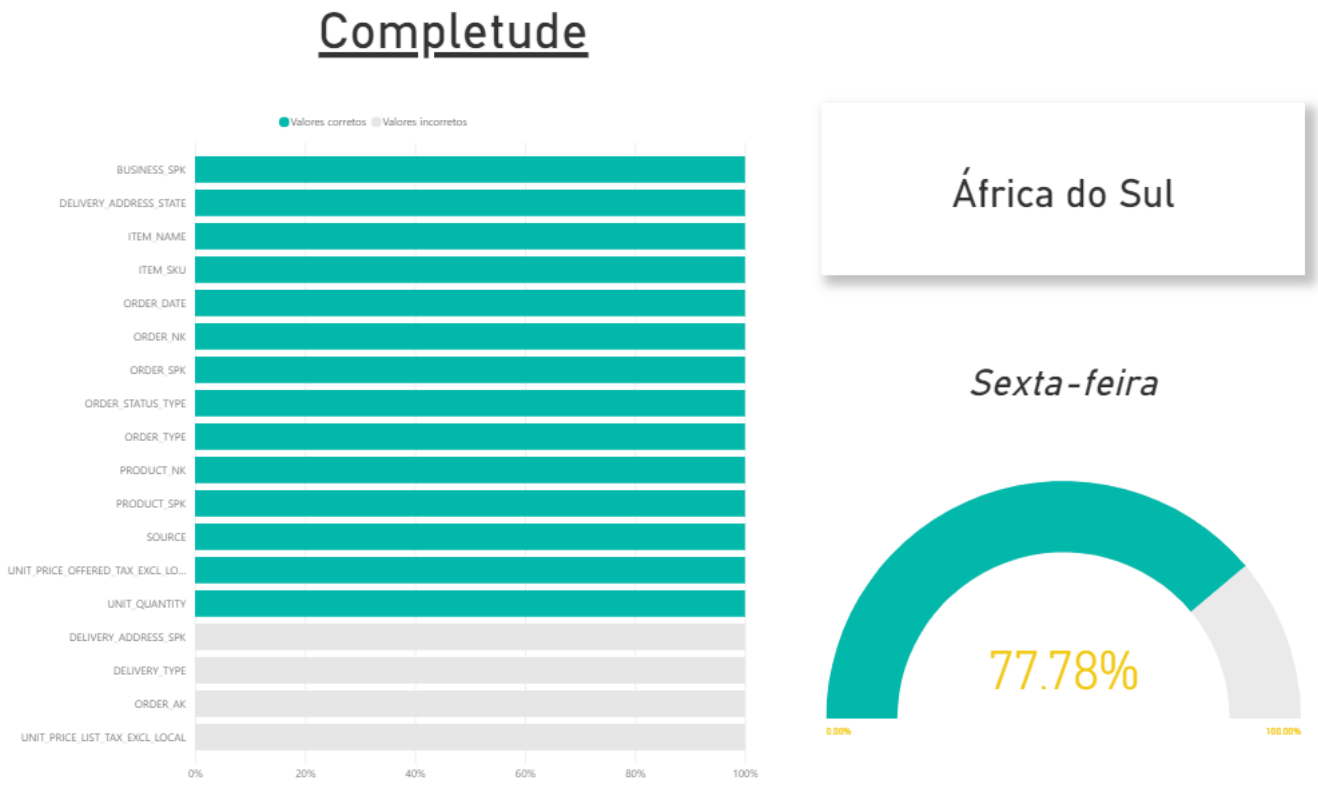
Figura 6.2: África do Sul - Completude, quarta-feira.



Já na quarta-feira, o percentual de completude da África do Sul subiu. Obteve-se 88,89% das suas colunas com valores positivos do ponto de vista de completude. Duas colunas se mantiveram com 0% de registros preenchidos, sendo elas *"ORDER AK"* e *"UNIT PRICE TAX EXCL LOCAL"*. As outras duas colunas que estiveram com problemas na segunda-feira, dessa vez, tiveram 100% de diferentes de *"null"*. Esse fato é o motivo do percentual geral de completude para a quarta-feira ter sido maior do que o valor de segunda-feira.

- Sexta-feira: 77,78 %

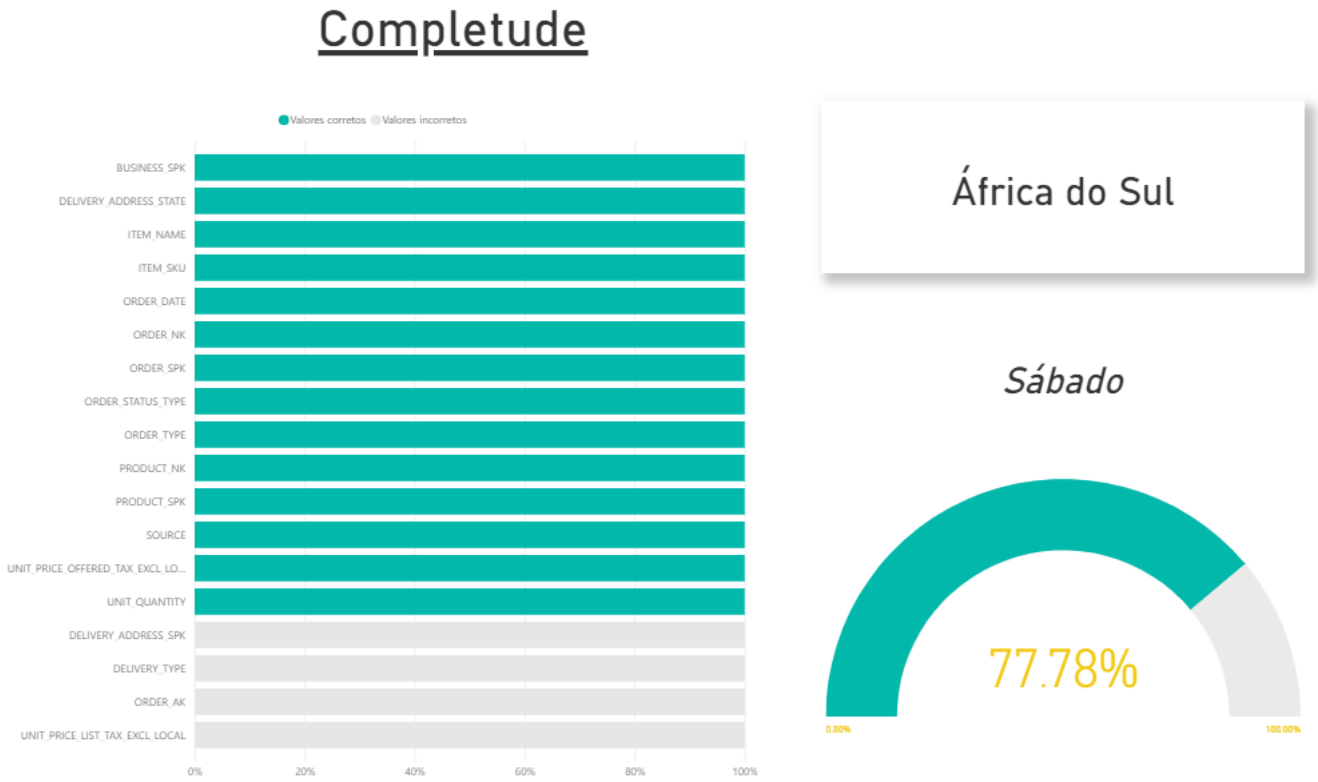
Figura 6.3: África do Sul - Completude, sexta-feira.



Na sexta-feira, o valor geral de completude voltou a ser 77,78%, com os exatos mesmos percentuais para cada uma das colunas. Assim como na segunda-feira, as colunas *"DELIVERY ADDRESS SPK"*, *"DELIVERY TYPE"*, *"ORDER AK"* e *"UNIT PRICE LIST TAX EXCL LOCAL"* tiveram seus registros com 0% de valores diferentes de "null".

- Sábado: 77,78 %

Figura 6.4: África do Sul - Completude, sábado.



Assim como na segunda-feira e na sexta-feira, o percentual de completude para as colunas do domínio de dados da África do Sul teve a média de 77,78%. Observa-se, novamente, que as colunas *"DELIVERY ADDRESS SPK"*, *"DELIVERY TYPE*, *"ORDER AK"* e *"UNIT PRICE LIST TAX EXCL LOCAL"* tiveram 0% de registros diferentes de *"null"*, ou seja, completamente vazios.

Os percentuais de completude gerais diários para os dias correspondentes a segunda-feira, quarta-feira, sexta-feira e sábado, foram, respectivamente: 77,78%, 88,89%, 77,78% e 77,78%.

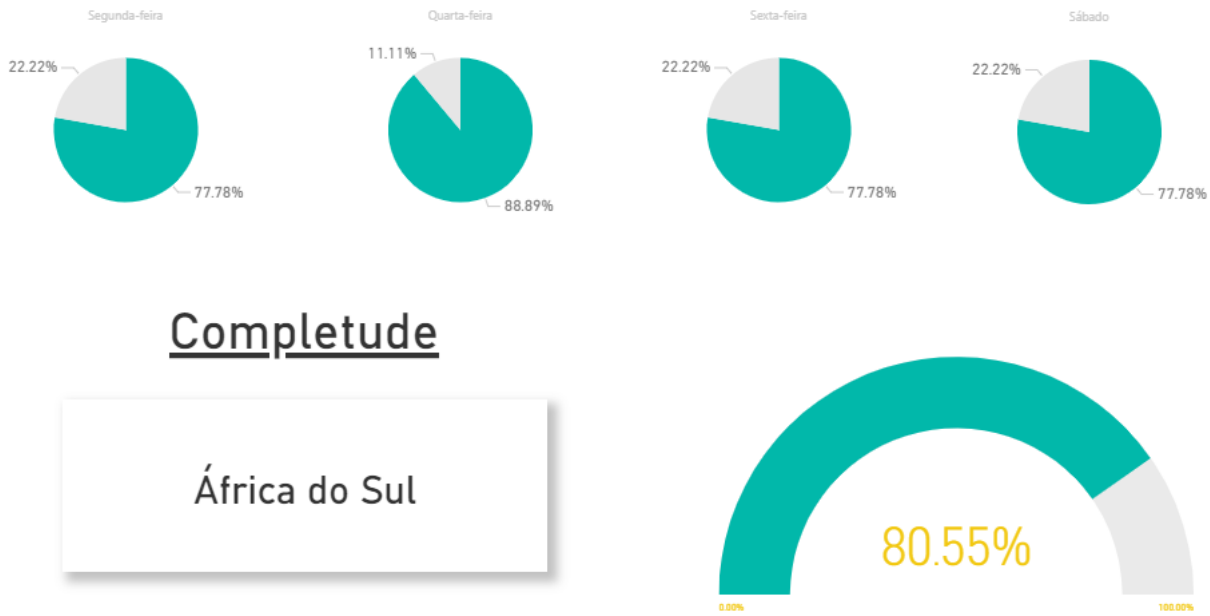
Com base nos percentuais obtidos, um padrão se torna evidente: em relação à completude, ou as colunas estão totalmente preenchidas com registros ou não possuem preenchimento algum. Este padrão é coerente, pois é possível levantar hipóteses como erros na captura da informação destes campos ou na integração de um destes campos específicos com a base de dados.

Das 18 colunas avaliadas em completude, 2 apresentaram 100% de valores incorretos, o que aponta para um problema claro de qualidade e invalidando sua utilização para análises de dados corporativas. Foram elas: *"ORDER AK"* e *"UNIT PRICE LIST TAX EXCL LOCAL"*.

Além disso, apenas outras 2 colunas apresentaram erros, sendo elas *"DELIVERY ADDRESS SPK"* e *"DELIVERY TYPE"*. Estas apresentam 0% de valores corretos no dia ou 100% corretos, o que aponta para problemas de registros conforme o dia em que são capturados.

O percentual final da África do Sul, sendo este a média dos percentuais dos quatro dias da semana em que foram capturados, foi de 80,55%. Este valor pode ser visualizado na figura 6.5.

Figura 6.5: África do Sul - % final geral de completude.



O resultado de 80,55% foi considerado satisfatório em razão da natureza da análise representar sempre colunas com 0% ou 100% de qualidade. Dessa forma, as informações contidas nas colunas corrompidas podem ser removidas das análises de dados corporativas, sem impactar a análise das demais informações do domínio de dados. Este fato também contribui para avaliar a conexão com estes atributos específicos, sabendo que, a depender do dia, possui registros finais válidos ou inválidos.

6.2 Análise de Conformidade

Na avaliação de conformidade, verificamos se os valores recebidos nas colunas estavam em conformidade com os padrões dos tipos de dados esperados. Ou seja, agora, além de se esperar receber dados em cada coluna, como em completude, espera-se receber dados no padrão correto. Para tal avaliação, foi utilizado como exemplo o domínio de

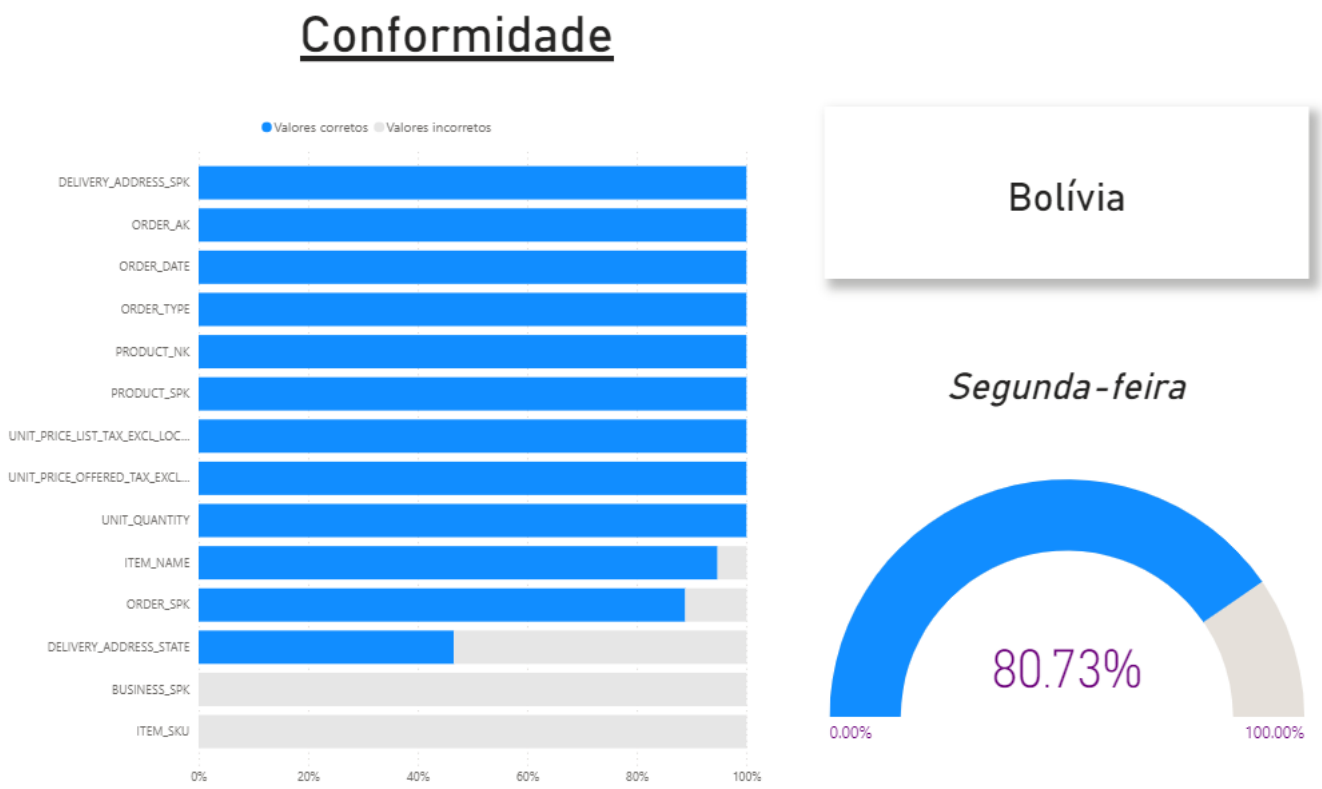
dados do país Bolívia.

A análise de conformidade rodou sobre 14 das 18 colunas do domínio de dados. Foram elas: *"BUSINESS SPK"*, *"DELIVERY ADDRESS SPK"*, *"DELIVERY ADDRESS STATE"*, *"ITEM NAME"*, *"ITEM SKU"*, *"ORDER AK"*, *"ORDER DATE"*, *"ORDER SPK"*, *"ORDER TYPE"*, *"PRODUCT NK"*, *"PRODUCT SPK"*, *"UNIT PRICE OFFERED TAX EXCL LOCAL"*, *"UNIT PRICE LIST TAX EXCL LOCAL"* e *"UNIT QUANTITY"*.

Espera-se 14 percentuais de qualidade de conformidade, um para cada coluna, bem como um percentual geral a cada dia, correspondendo à média dos percentuais das colunas do dia em questão. Ao final, espera-se um percentual final geral de qualidade, representando a média geral dos 4 dias da semana de captura de dados. Para cada dia em que os dados foram capturados, os resultados obtidos foram os seguintes:

- Segunda-feira: 80,73%

Figura 6.6: Bolívia - Conformidade, segunda-feira.

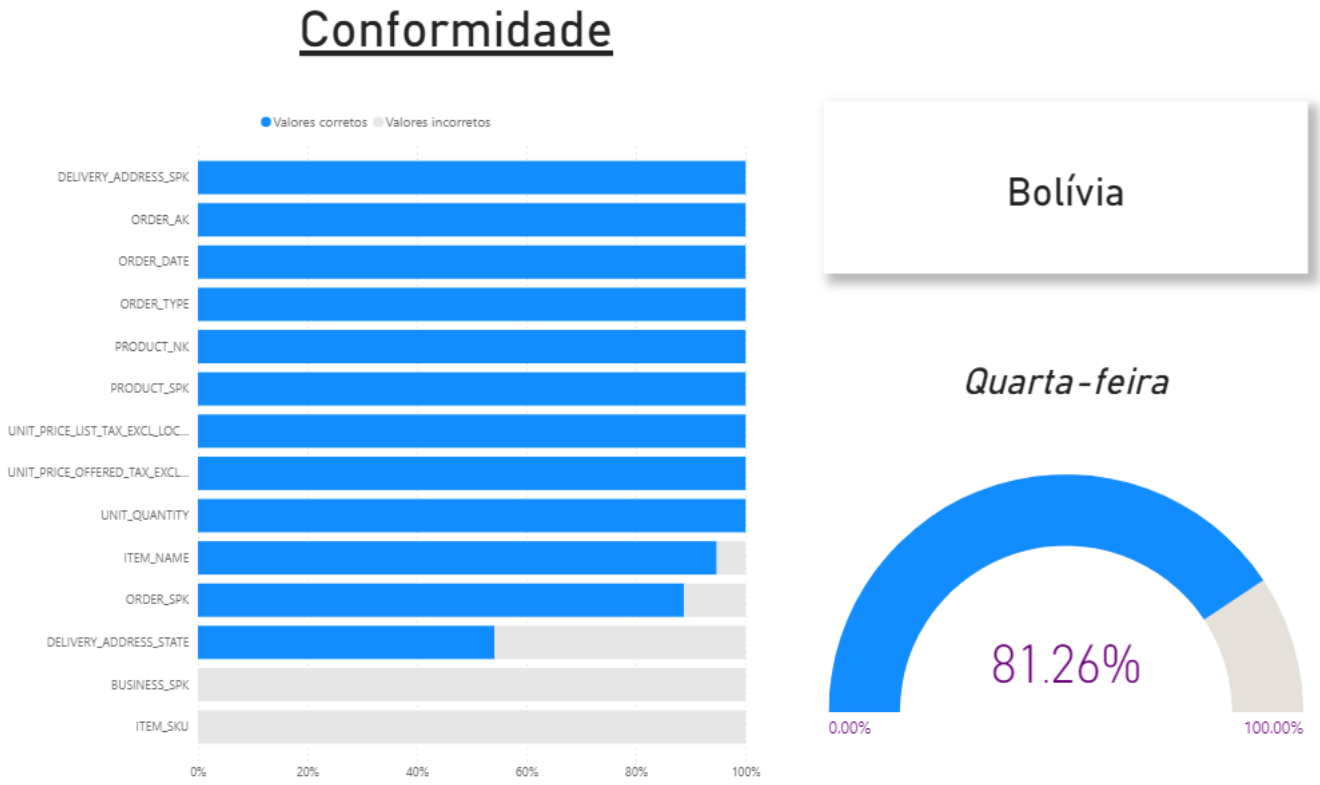


Na segunda-feira, a Bolívia teve 80,73% dos registros do seu domínio de dados em conformidade com os padrões esperados. Nota-se que 9 das 14 colunas tiveram 100% de qualidade nos seus valores. As colunas *"BUSINESS SPK"* e *"ITEM SKU"* tiveram 0% de qualidade, enquanto a coluna *"ITEM NAME"*, obteve 94,73% de valores em conformidade, a coluna *"DELIVERY ADDRESS STATE"* obteve 46,59%

e a coluna *"ORDER SPK"* ficou com 88,83%.

- Quarta-feira: 81,26%

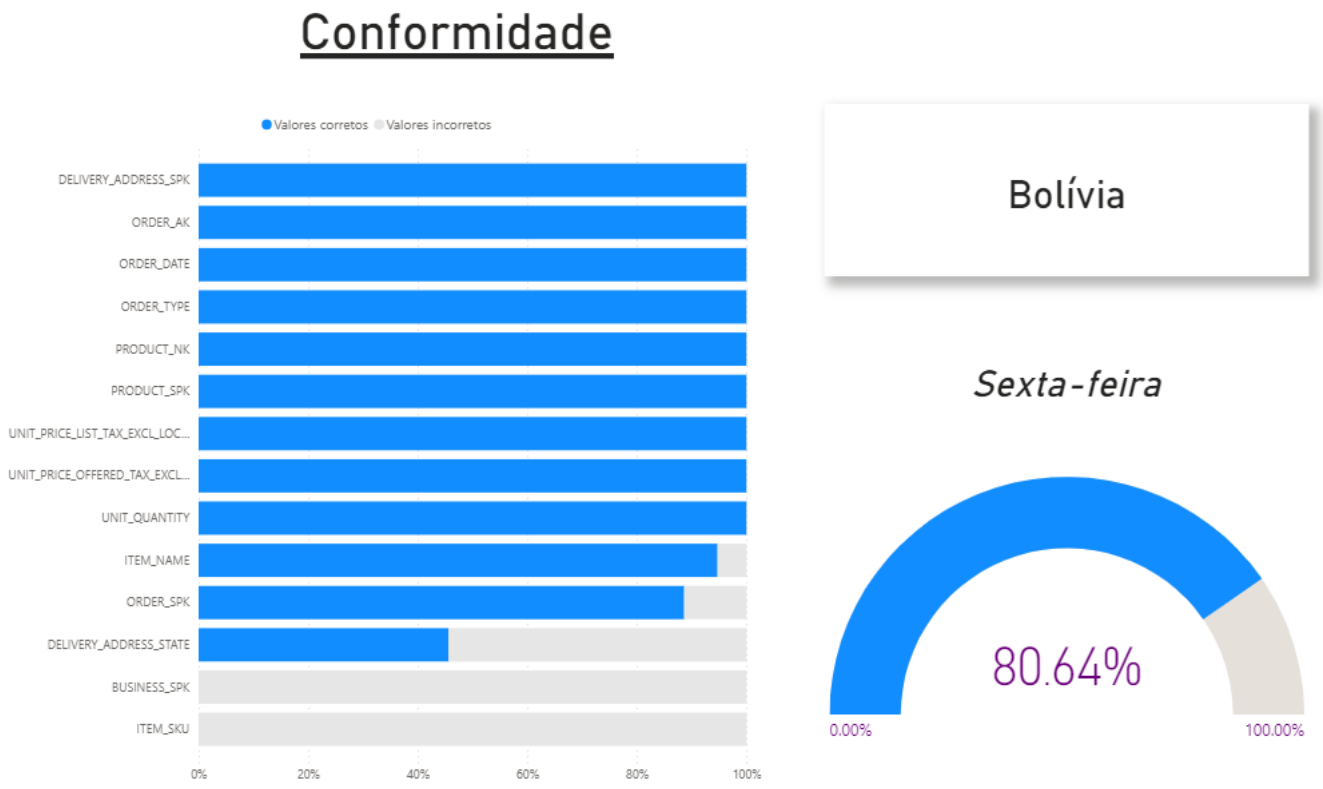
Figura 6.7: Bolívia - Conformidade, quarta-feira.



O percentual de conformidade dos dados da Bolívia na quarta-feira foi de 81,26%, muito próximo do valor apresentado na segunda-feira. Os valores de qualidade das colunas também foram parecidos: 9 das 14 colunas tiveram 100% de qualidade nos seus valores, com as colunas *"BUSINESS SPK"* e *"ITEM SKU"* com 0% de qualidade, enquanto a coluna *"ITEM NAME* obteve 94,73% de valores em conformidade, a coluna *"DELIVERY ADDRESS STATE"* obteve 54,18% e a coluna *"ORDER SPK"* com 88,77%.

- Sexta-feira: 80,64%

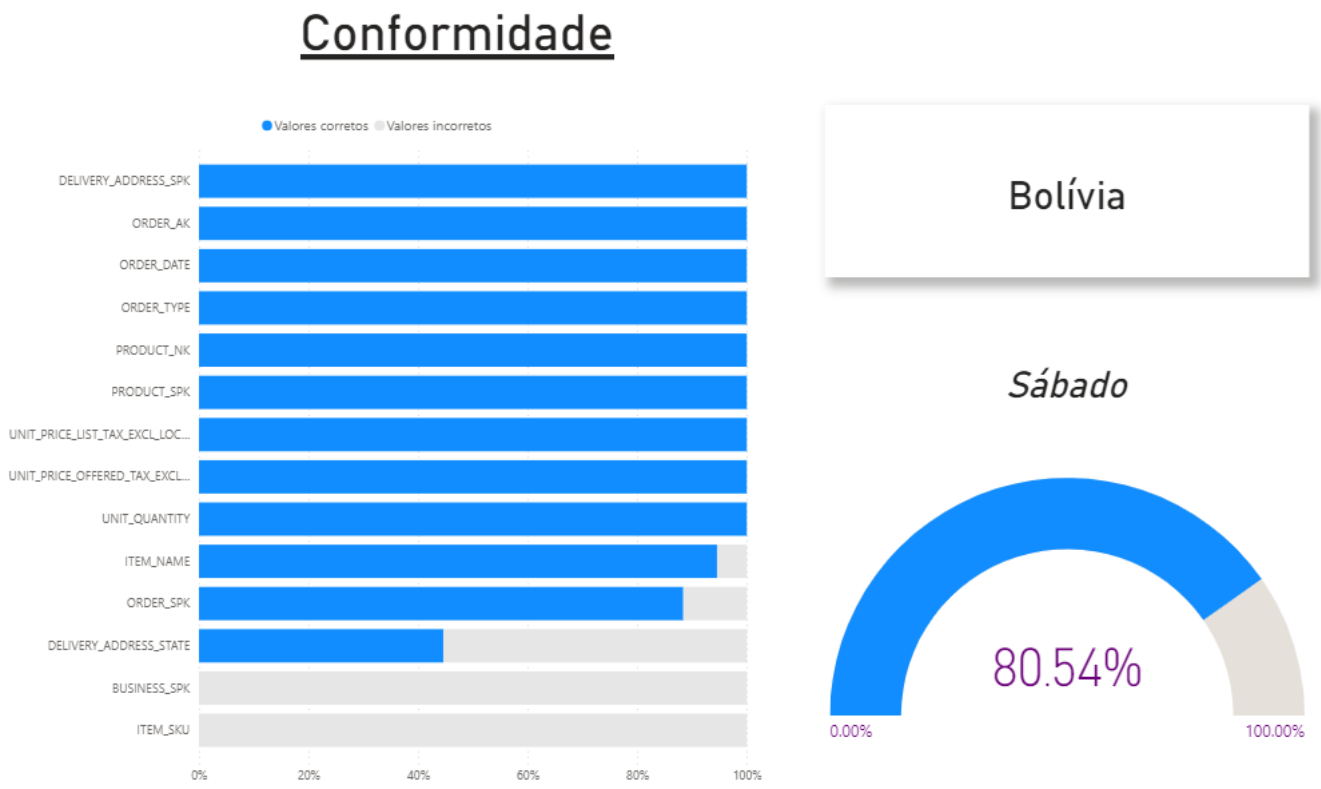
Figura 6.8: Bolívia - Conformidade, sexta-feira.



Na sexta-feira, a Bolívia manteve um percentual de conformidade muito próximo dos dias anteriores, com, agora, 80,64% dos seus registros com qualidade. Novamente, apenas 5 colunas não tiveram seus valores com o percentual de 100% de qualidade. As colunas *"BUSINESS SPK"* e *"ITEM SKU"* tiveram 0% de valores coerentes, enquanto a coluna *"ITEM NAME"* obteve 94,73%, a coluna *"DELIVERY ADDRESS STATE"* obteve 45,63% e a coluna *"ORDER SPK"* ficou com 88,64%.

- Sábado: 80,54%

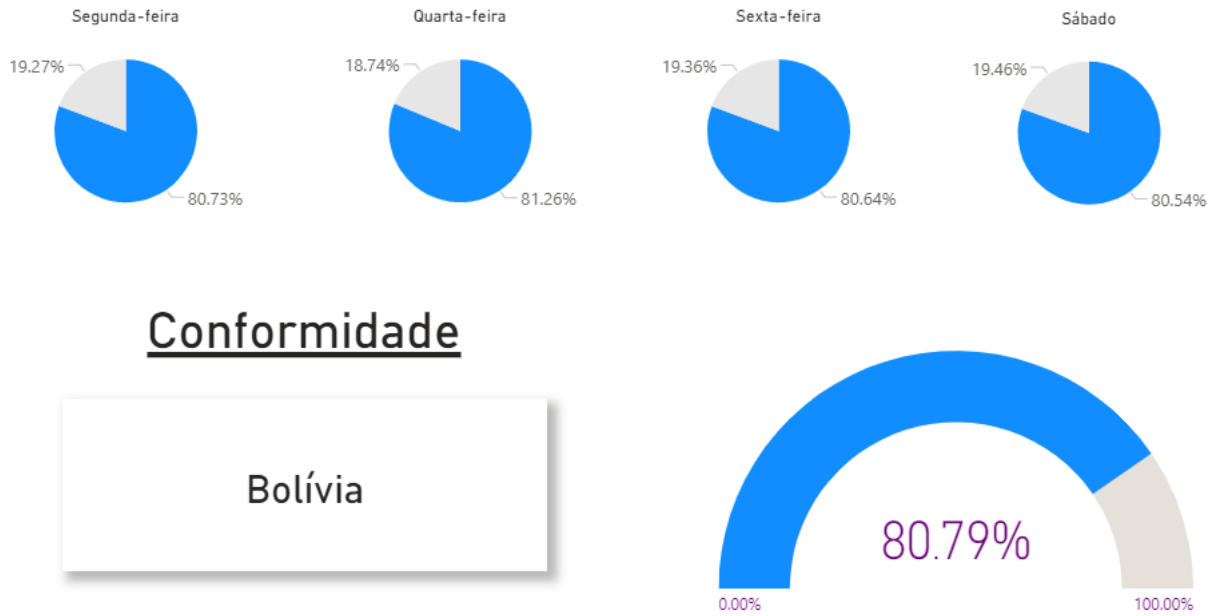
Figura 6.9: Bolívia - Conformidade, sábado.



No sábado, embora o percentual tenha se mantido, em todos os dias, sempre muito semelhante, foi observado o menor dos percentuais da semana com 80,54% de registros com qualidade no quesito conformidade. Assim como nos outros dias de captura de dados, as colunas *"BUSINESS SPK"* e *"ITEM SKU"* tiveram 0% de registros coerentes. As colunas *"ITEM NAME"*, *"DELIVERY ADDRESS STATE"* e *"ODER SPK"* tiveram, respectivamente, 94,57%, 44,60% e 88,41% de registros em conformidade. As outras 9 colunas tiveram seus registros com 100% de qualidade.

Como percentual geral final de conformidade, considerando todos os dias da semana em que foram capturados dados de análise, a Bolívia teve 80,79% de registros em conformidade. Este resultado não foi considerado satisfatório, uma vez que os percentuais das colunas variaram pouco ao longo dos dias, o que aponta para possíveis problemas na elaboração das expressões regulares de tipos de valores esperados para as colunas que apresentaram problemas. Para ter certeza dessa possível inconsistência, faz-se necessária uma avaliação manual dos registros incorretos de cada dia, que não foi contemplada neste trabalho.

Figura 6.10: Bolívia - % final geral de conformidade.



É relevante destacar que os valores de registros incorretos não são resultantes da ausência de valores ou de valores *"null"*. Porém, as colunas *"ORDER AK"* e *"UNIT PRICE LIST TAX EXCL LOCAL"*, que tiveram seus valores com 0% de completude em todos os dias de análise, ou seja, que agora não possuíam registros para medir conformidade, acabaram apresentando sempre 100% de qualidade em conformidade, o que sabe-se que está incorreto. Também, a coluna *"DELIVERY ADDRESS SPK"*, que, com 100% ou 0% de completude, apresentou 100% de conformidade. Essas inconsistências apresentam falhas na captura da medida de qualidade de conformidade.

6.3 Análise de Precisão

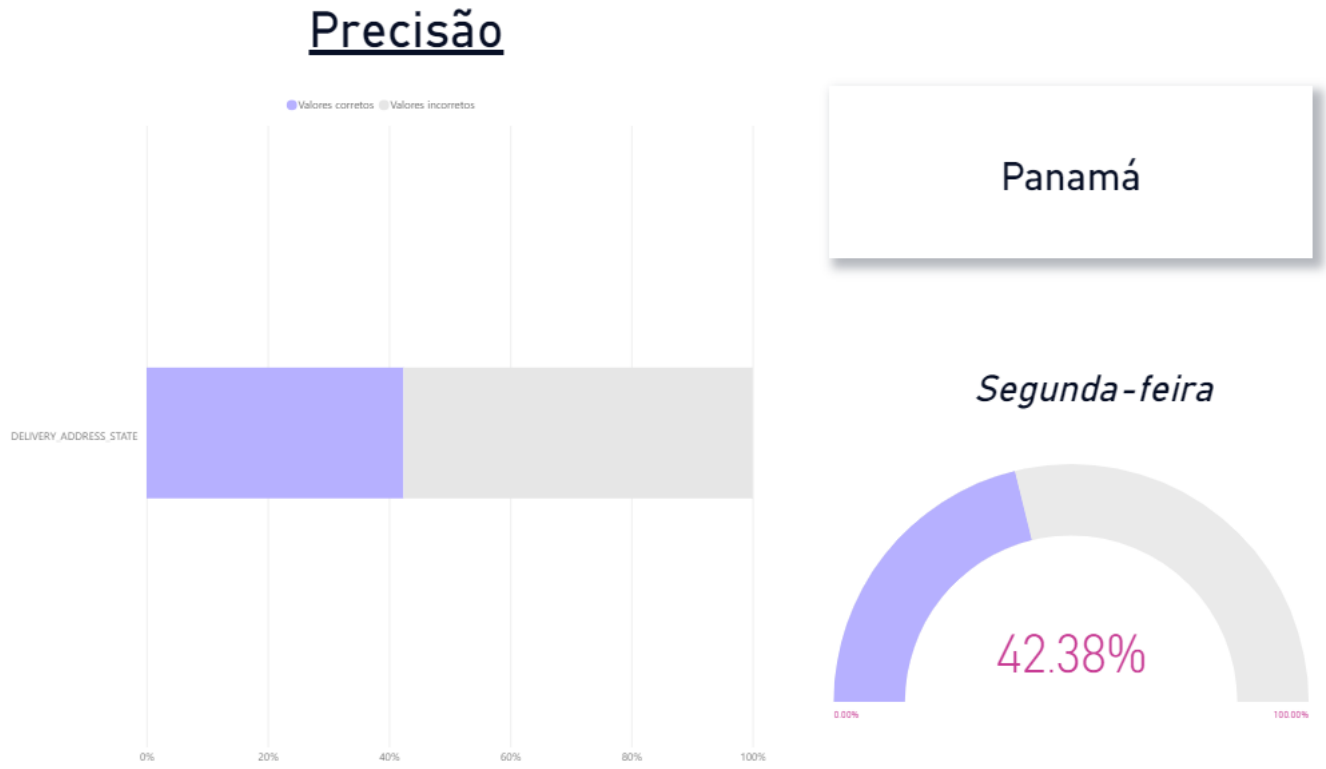
Na análise de precisão, foi escolhido o país Panamá para ter seus dados utilizados como exemplo de análise. Nessa avaliação, a verificação se dá sobre a comparação dos valores recebidos em cada registro com uma lista previamente definida de valores válidos. A coluna utilizada para esta avaliação é a *"DELIVERY ADDRESS STATE"*, que corresponde ao registro que guarda a informação do estado do país em que a compra do pedido foi realizada. Espera-se, então, um percentual de qualidade de precisão para a coluna em questão a cada dia, que corresponde ao percentual de qualidade de precisão geral deste dia, já que apenas 1 coluna foi avaliada com esta métrica.

Ao final, conta-se com um percentual geral final do país para a dimensão de pre-

cisão, que corresponde à média dos percentuais dos 4 dias de avaliação. Para cada dia em que os dados foram capturados, obtivemos os seguintes resultados:

- Segunda-feira: 42,38%

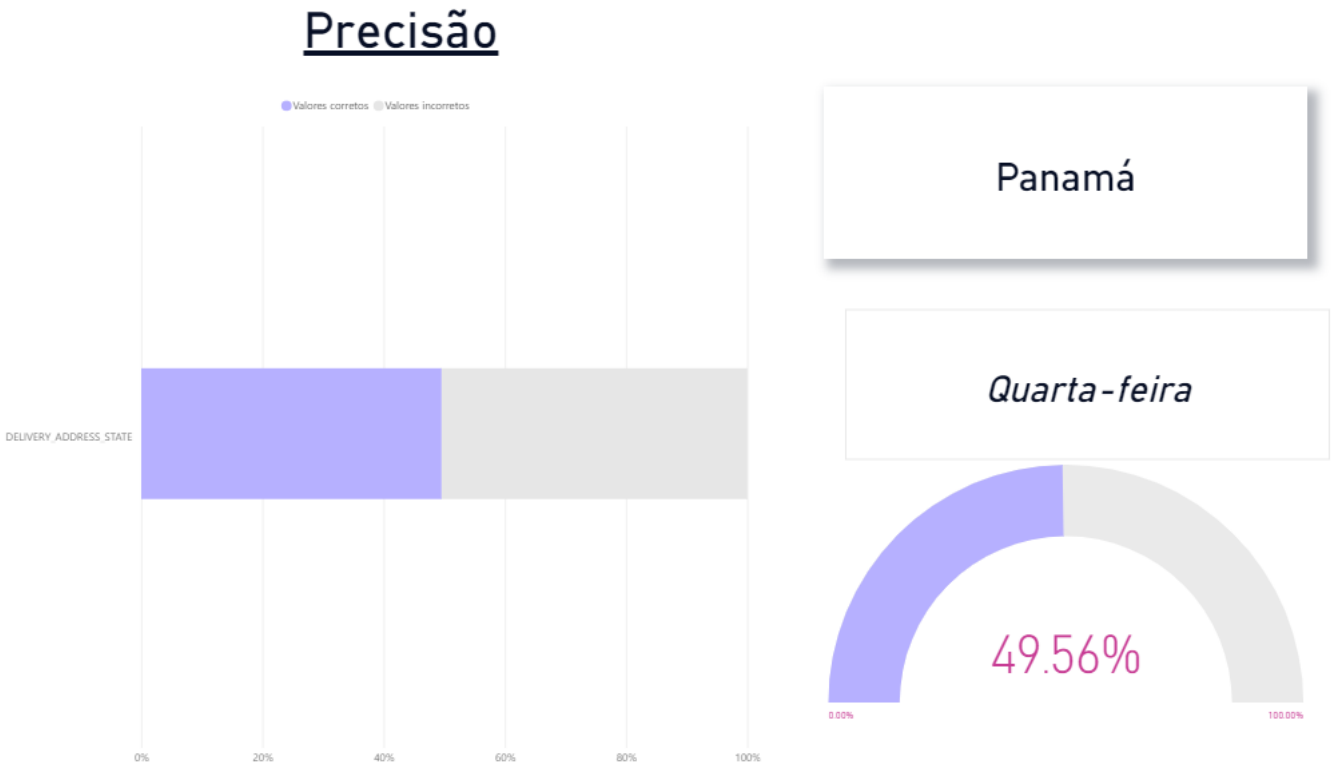
Figura 6.11: Panamá - Precisão, segunda-feira.



Na segunda-feira, o Panamá contou com 42,38% dos seus registros para a informação de estado com valores corretos.

- Quarta-feira: 49,56%

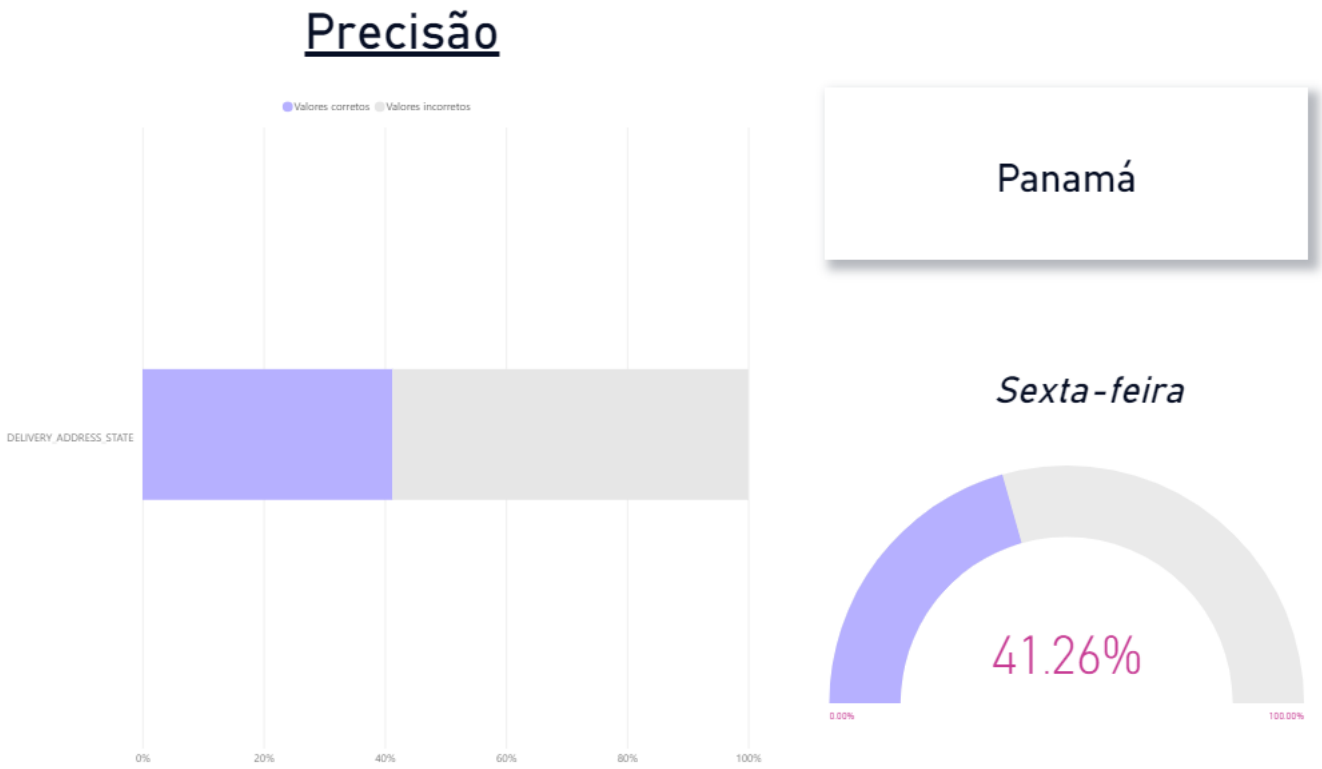
Figura 6.12: Panamá - Precisão, quarta-feira.



Na quarta-feira, o percentual de precisão dos seus registros para a informação de estado foi de 49,56%.

- Sexta-feira: 41,26%

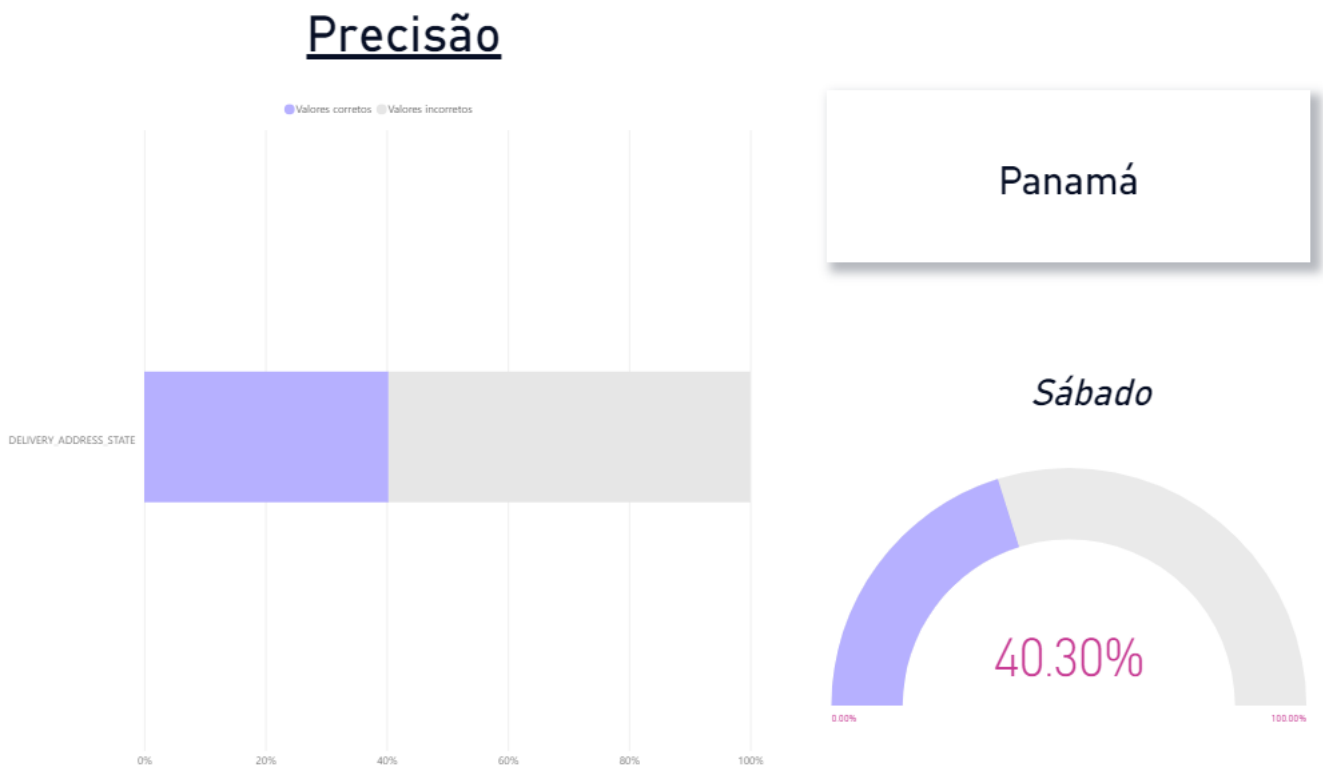
Figura 6.13: Panamá - Precisão, sexta-feira.



Na sexta-feira, o percentual de precisão dos dados se manteve, totalizando 41,26% dos seus registros com informações válidas.

- Sábado: 40,30%

Figura 6.14: Panamá - Precisão, sábado.



Já no sábado, foi capturado o percentual de 40,30% de qualidade dos dados no quesito precisão.

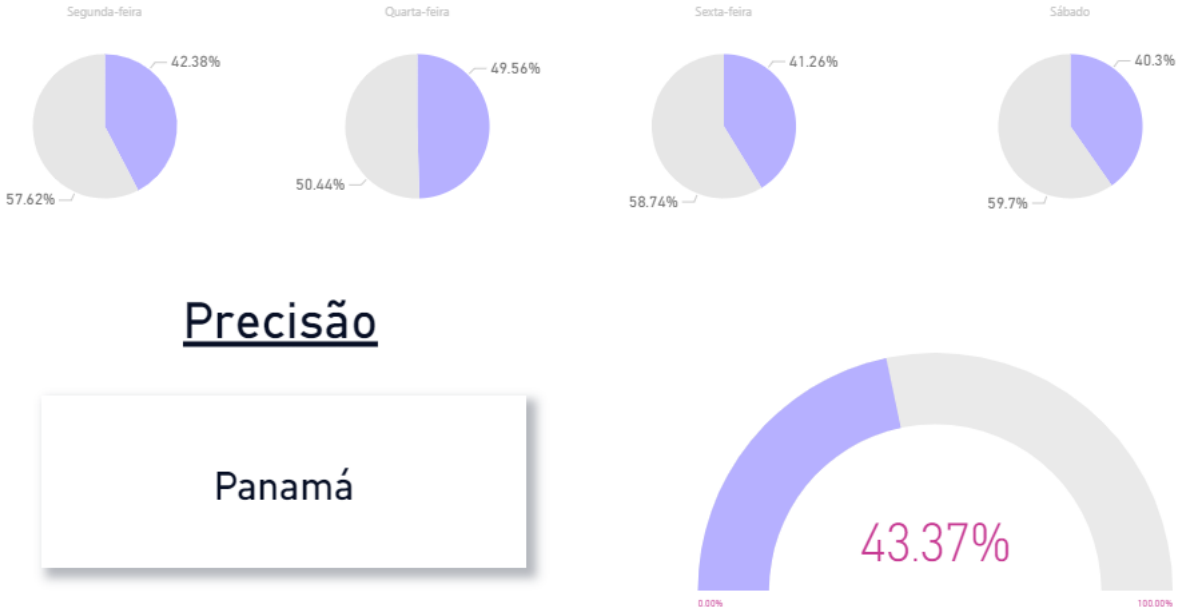
Ao considerarmos a análise de precisão, observamos que registros com erros podem coincidir com registros que não estão em conformidade. De fato, parece haver uma sobreposição desses problemas.

Tomando como exemplo a sexta-feira, a coluna que contém o estado de entrega (*"Delivery address state"*) apresenta 41,26% de valores válidos em termos de precisão. Ao mesmo tempo, esta coluna, no mesmo dia, possui 45,63% de registros válidos em relação à conformidade. Essa observação sugere que os registros 45,63% dos registros de estado deste dia correspondem a palavras, mas somente 41,26% são, de fato, estados válidos do país. Os 4,37% de registros de diferença entre os valores apontam para a possibilidade de serem palavras, mas não precisas em relação a valores de estados do país.

Os resultados obtidos para precisão foram considerados coerentes, mas não satisfatórios. Com apenas 42,38% de valores com qualidade em precisão, não torna-se possível utilizar esse campo de informações em análises de dados corporativas. Porém, sabendo-se que a maior parte dos erros de precisão correspondem a erros de conformi-

dade, é possível rever as formas de captura e armazenamento dessa informação, onde, corrigindo a conformidade dos registros, é provável que os valores de precisão cresçam bastante.

Figura 6.15: Panamá - % final geral de precisão.



7 CONCLUSÃO

O presente trabalho se dedicou à análise de três dimensões de qualidade de dados aplicadas ao domínio de uma empresa brasileira do ramo varejista, sendo as dimensões: completude, conformidade e precisão. O objetivo principal consistiu em calcular os percentuais de qualidade para cada uma dessas dimensões, a fim de identificar os dados aptos para análises corporativas e compreender os tipos de problemas associados a dados incorretos nas dimensões analisadas.

Para avaliar a completude, verificou-se se os registros das colunas selecionadas continham dados não nulos. A conformidade foi analisada por meio de expressões regulares, que definiram os tipos de dados aceitáveis para as colunas especificadas. A medição de precisão envolveu uma lista predefinida de valores corretos.

Para alcançar esses objetivos, coletaram-se dados de onze países diferentes, durante quatro dias, do domínio da empresa, sendo eles: África do Sul, Argentina, México, Bolívia, Paraguai, Colombia, Panamá, República Dominicana, Honduras, El Salvador e Peru. Um código em *python* foi desenvolvido para a leitura e análise da qualidade dos dados, com o suporte da biblioteca *Great Expectations*. O *software PowerBI* foi utilizado para apresentar os resultados e calcular os percentuais de qualidade.

Cada país e dimensão de qualidade foram detalhadamente abordados, apresentando os percentuais de registros de qualidade para cada coluna de dados. A cada dia, demonstrou-se o percentual geral de registros de qualidade para cada dimensão. Ao final, foram apresentados os percentuais gerais de qualidade de dados para cada dimensão, considerando os dados de quatro dias para cada país.

Como perspectivas futuras, recomenda-se ampliar o número de dias de coleta e análise de dados, permitindo maior embasamento para a inferência de possíveis causas dos problemas de qualidade. Esta é uma das recomendações que são indicadas em "*Data Quality for Analytics Using SAS*", de Gerhard Svolba e Ina Felsheim, e que podem ser positivas no processo de evolução de trabalhos em qualidade de dados.

Além disso, para análise de conformidade, sugere-se uma definição mais aprofundada das expressões regulares para identificar com maior precisão os tipos de dados esperados em cada coluna. Também, faz-se necessária uma análise mais aprofundada sobre os resultados que apresentaram dados não completos, mas conformes. Estes precisam ser reavaliados para não apresentar qualidade de conformidade quando não há dados para avaliar.

Além disso, para a análise de precisão, é interessante verificar os dados que estão conformes, mas não precisos, a fim de verificar se estes são dados que estão próximos dos valores pré definidos e que poderiam ser melhor escritos na lista prévia.

8 REFERÊNCIAS

English, L. P. (1999). Information Quality: The Potential of Data and Analytics to Generate Knowledge.

Inmon, W. H. (2005). Building the Data Warehouse.

Olson, J. E. (2003). Data Quality: The Accuracy Dimension. Morgan Kaufmann.

Maydanchik, A. (2007). Data Quality Assessment.

English, L. P. (2009). Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits.

Batini, C., Scannapieco, M., & Data Quality: Concepts, Methodologies and Techniques. (2006).

McGilvray, D. (2008). Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information.

Redman, T. C. (2008). Data Driven: Profiting from Your Most Important Business Asset.

Kimball, R. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.

Redman, T. (2001). Data Quality: A Survival Guide.

Svolba, G., & Felsheim, I. (2015). Data Quality for Analytics Using SAS.

Wixom, B. H., Ariyachandra, T., Douglas, D. E., Goul, M., Gupta, B., Iyer, L. S., ... & Turetken, O. (2014). The current state of business intelligence in academia: The arrival of big data. Communications of the Association for Information Systems, 34(1), 1-41.

Kimball, R., & Ross, M. (2002). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Wiley.

Kimball, R., & Ross, M. (2002). The Data Warehouse Lifecycle Toolkit. Wiley.

Inmon, W. H., & Hackathorn, R. D. (1994). Using the Data Warehouse. Computing McGraw-Hill.