Escopo de Negócio

Adventure Works

1. Introdução

A Adventure Works (AW), uma indústria de bicicletas em franco crescimento, busca utilizar seus dados de forma estratégica para guiar decisões e se diferenciar da concorrência.

Com mais de 500 produtos, 20.000 clientes e 31.000 pedidos, a empresa vê no uso de dados uma oportunidade de se tornar "data driven" e otimizar suas operações, especialmente na área de vendas, que será o foco inicial do projeto.

Este relatório detalha o escopo, os objetivos e os passos necessários para a implementação de uma infraestrutura moderna de dados para analytics e a criação de um modelo preditivo para prever a demanda dos próximos três meses por produto e loja.

2. Objetivos

O projeto tem como principais objetivos:

- Desenvolver uma infraestrutura robusta de analytics que permita a centralização e o cruzamento de dados de diferentes sistemas utilizados pela AW, começando pela área de vendas.
- Implementar um data warehouse (DW) modelado com tabelas de fatos e dimensões para responder às perguntas de negócio estratégicas, utilizando o Snowflake como base de dados e o dbt para transformação de dados.
- Criar um dashboard interativo em Power BI para visualização dos principais indicadores de vendas, fornecendo insights estratégicos para a diretoria e equipe operacional.
- Desenvolver um modelo preditivo de demanda, utilizando técnicas de machine learning, para prever a demanda de produtos nos próximos 3 meses em cada loja.
- Demonstrar o valor da infraestrutura e dos modelos preditivos à diretoria e outras áreas, como Planejamento de Demanda, gerando análises que superem as atuais previsões feitas com médias móveis no Excel.

3. Recursos Disponíveis

Pessoas envolvidas no projeto:

Nome	Função	Empresa
Carlos Silveira	CEO	AW
João Muller	Diretor de Inovação	AW
Silvana Teixeira	Diretora Comercial	AW
Nilson Ramos	Diretor de TI	AW
Luís Soares	Gestor de Planejamento	AW
Gabriel Santos	Analista	AW
Drielly Sanches	Cientista de Dados	Indicium

4. Escopo do Projeto

O que o projeto vai atender

1. **Implementação de um DW**: O desenvolvimento da modelagem dimensional (fatos e dimensões) será o primeiro passo, abrangendo principalmente a área de vendas (orders), com potencial de expansão para outras áreas no futuro.

2. **Transformação de dados com dbt**: Uso de dbt para realizar as transformações necessárias e preparar os dados para as análises.

3. **Criação de dashboards**: Desenvolvimento de dashboards em Power BI para visualização e monitoramento de KPIs críticos de vendas.

4. **Modelo preditivo de demanda**: Criação de um modelo de previsão de demanda utilizando técnicas como séries temporais hierárquicas, treinado com dados históricos de vendas, para prever a demanda de produtos por loja.

O que o projeto não vai atender

Para garantir um foco claro e direcionado, o projeto excluirá algumas áreas que não são essenciais para esta análise:

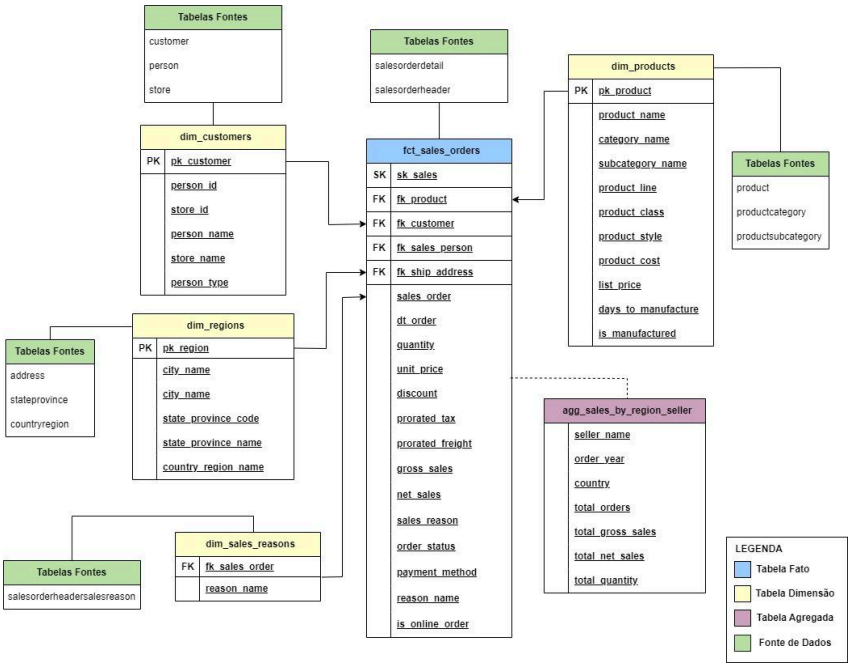
1. **Integração completa com todos os sistemas**: Em um primeiro momento, o foco será na área de vendas, com a integração de outras áreas (como marketing ou finanças) sendo planejada para fases futuras.

5. Metodologia

A metodologia adotada para este projeto segue uma abordagem estruturada, garantindo que todas as etapas sejam realizadas de forma eficiente. Cada fase do projeto foi desenhada para maximizar a qualidade das entregas e facilitar a adoção da infraestrutura de dados e dos modelos analíticos pela Adventure Works.

5.1. Modelagem Dimensional (DW):

No processo de modelagem dimensional do Data Warehouse, foi feita uma exploração inicial dos dados disponibilizados pela Adventure Works de forma pública com todos os schemas e fontes de dados dos seus departamentos empresariais. A partir desse estudo, foi desenhado o Diagrama Conceitual abaixo com a estruturação das dimensões e fatos que seriam importantes para a construção do BI estratégico e operacional da empresa.



Dessa forma, o DW foi configurado com as seguintes tabelas dimensões, fatos e agregadas:

- **Dimensões:**
  - dim_customers: dados sobre os clientes.
  - dim_products: dados sobre os produtos.
  - dim_regions: dados de localização.

    - dim_sales_reason: dados sobre os motivos de compra.
- **Fato:**
    - fct_sales_orders: dados sobre as ordens de compra por cliente e por produto com métricas de preço, quantidade, venda bruta, venda líquida, entre outros.
- **Agregada:**
    - agg_sales_by_region_seller: dados agregados de vendas por região e por vendedor.

Com base nessa modelagem, foi realizada a implementação da infraestrutura moderna de dados com a utilização do Snowflake como plataforma de configuração do Data Warehouse.

O Snowflake possui uma arquitetura baseada em nuvem que facilita a integração com outras plataformas, oferecendo desempenho otimizado com custos controlados. Além disso, ele suporta múltiplos tipos de dados e permite a consolidação de fontes de dados distintas, fornecendo segurança robusta, recuperação de falhas automática e colaboração simplificada entre equipes e parceiros de negócio. Por conta disso, essa ferramenta foi escolhida entre as demais oferecidas pelo mercado.

5.2. Transformação e Preparação dos Dados:

Os dados foram transformados utilizando o dbt cloud, sendo tratados e preparados para a visualização e análises preditivas. Essa fase incluiu a limpeza de dados e a criação de métricas para os dashboards.

Por meio do dbt cloud, foram criadas três camadas de transformação de dados, sendo:

1. **Staging:** Camada inicial onde os dados brutos são carregados e padronizados para facilitar o uso nas etapas seguintes. Aqui, os dados foram apenas renomeados e reorganizados.
2. **Aggregate:** Camada onde os dados já limpos são organizados em tabelas que atendem a necessidades de áreas de negócio específicas, como vendas ou marketing, facilitando o acesso a informações mais refinadas e agregadas. Essas tabelas combinam métricas de vários processos em um nível comum de detalhe, otimizando o acesso aos dados.
3. **Marts:** Camada final de agregação, onde os dados são resumidos, calculados e consolidados em métricas e indicadores chave, prontos para visualizações e análises preditivas nos dashboards.

Os modelos da camada staging e aggregate foram materializados como *views*, enquanto os modelos da camada marts foram materializados como *tables*.

As *views* são consultas dinâmicas que não armazenam dados fisicamente. Elas são usadas para garantir que os dados dessas camadas estejam sempre atualizados, refletindo as mudanças nos dados brutos sem necessidade de armazenar informações redundantes.

Na camada staging, isso é importante, pois os dados estão sendo preparados e transformados de forma contínua. Na camada aggregate, as métricas são recalculadas dinamicamente sempre que necessário, o que mantém os resultados atualizados sem precisar reprocessar grandes volumes de dados.

As *tables* armazenam dados fisicamente, o que melhora o desempenho para consultas frequentes e repetitivas. Na camada *marts*, os dados já foram limpos e organizados para atender a necessidades específicas de áreas de negócio. Como esses dados são consultados regularmente, armazená-los como *tables* evita a necessidade de recalcular ou reprocessar as consultas a cada vez, melhorando a velocidade e eficiência das consultas.

Além disso, foram criados arquivos de configuração YAML para cada modelo, contendo definições claras sobre as tabelas e colunas, além de metadados e documentação.

Esses arquivos permitem o gerenciamento eficiente dos modelos no dbt, facilitando a organização, rastreabilidade e manutenção das transformações de dados. Neles, foram implementados testes de integridade como *unique* e *not null* para garantir a qualidade dos dados. O teste *unique* assegura que os valores de determinadas colunas não sejam duplicados, enquanto o teste *not null* verifica que os campos essenciais não contenham valores nulos, garantindo consistência e precisão nos dados.

5.3. Criação de Dashboards em Power BI:

Com os dados bem estruturados no DW, foi então iniciado o processo de *Business Intelligence* para construção de dashboards interativos para monitoramento de KPIs de vendas, como receita, quantidade de produtos vendidos, ticket médio, total de clientes, volume de vendas por loja, categoria de produto, entre outros. Foi também realizada a inclusão de filtros, gráficos de tendência e comparações de desempenho.

A ferramenta escolhida para esta etapa foi o Power BI, devido à sua facilidade de uso, integração com diversas fontes de dados, recursos avançados de visualização interativa e capacidade de criar dashboards dinâmicos que permitem a análise rápida e acessível de informações, facilitando a tomada de decisões estratégicas.

Os dashboards foram segmentados em:

- **Visão Estratégica:** dashboard com os principais indicadores a serem vistos diariamente pela gerência da empresa.

- **Visão Operacional:** dashboard mais granular, voltado para a operação, que permita a análise detalhada das vendas em diferentes dimensões:
    - Por pedido: Analisar periodicidade, ticket médio, quantidade, métodos de pagamento e motivos de compra.
    - Por região: Identificar regiões com maior participação de mercado.
    - Por cliente: Auxiliar em campanhas de e-mail marketing, destacando clientes mais frequentes e os que estão inativos.

5.4. Desenvolvimento do Modelo Preditivo:

A ferramenta de implementação utilizada nessa etapa foi o Google Colab e o *notebook* com todos os códigos e scripts pode ser acessado na seguinte página: Adventure Works Sales Forecasting.

Para importação dos dados relevantes para a análise, foi feita uma conexão com o Snowflake através das credenciais de acesso do DW. Em seguida, uma consulta SQL foi executada para extrair os dados, que foram armazenados em um dataframe do Pandas para posterior processamento e análise. As variáveis selecionadas para compor esse *dataframe* foram:

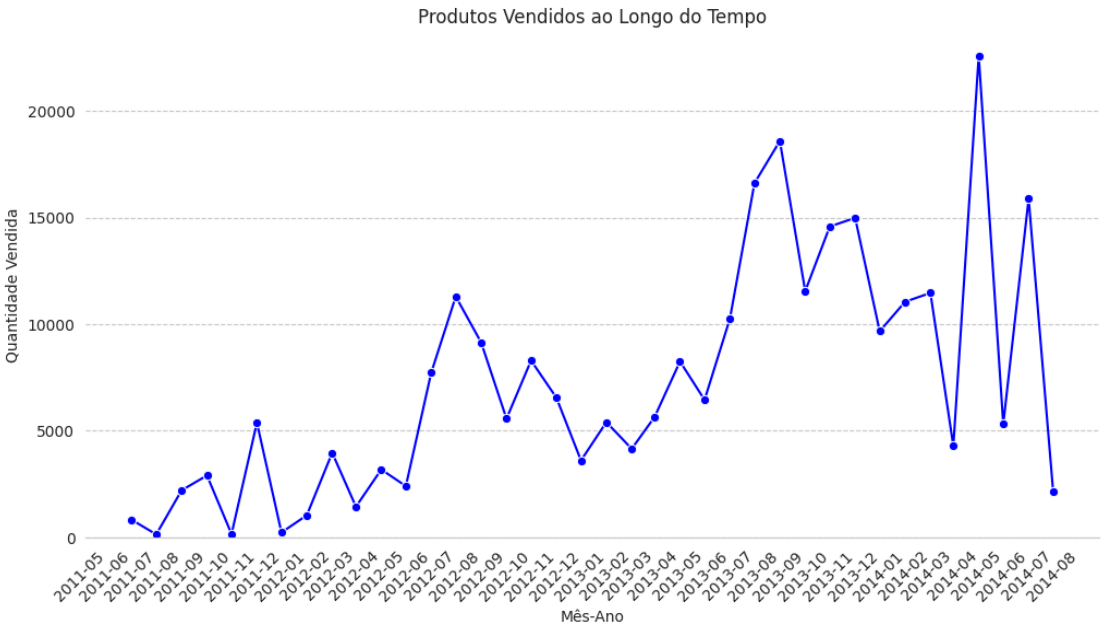
Coluna	Descrição
order_id	ID do pedido
dt_order	Data da compra
store_name	Nome da loja onde ocorreu a transação
category_name	Categoria do produto comprado
subcategory_name	Subcategoria do produto comprado
product_name	Nome do produto
quantity	Quantidade comprada do produto
is_online_order	Flag que indica se a compra foi online ou na loja
order_reason	Motivo da compra
state_province	Estado ou Província do Cliente
country	País do Cliente

Para realizar a previsão sobre a demanda dos próximos 3 meses de cada produto em cada loja da empresa, foi inicialmente realizada uma exploração dos dados e um estudo para verificar a presença ou não de sazonalidade nas vendas dos produtos.

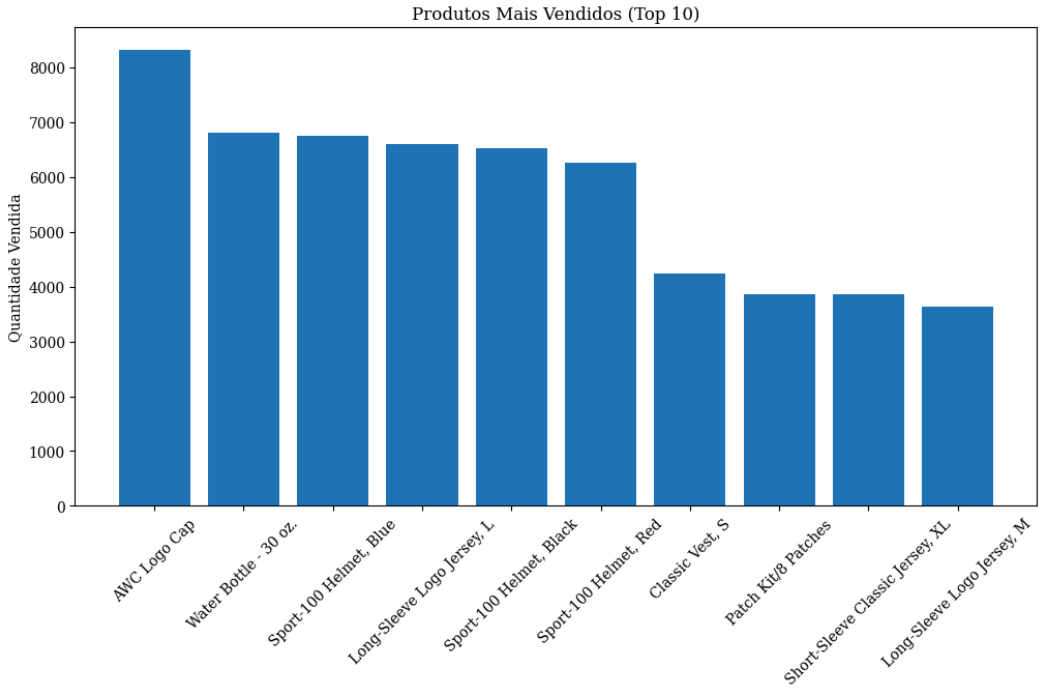
A partir desse estudo, foi verificada a presença de 60.398 valores nulos no campo *store_name*, que na verdade se tratavam das compras realizadas online. Portanto, esses valores nulos foram substituídos por “Online Store”.

Análise de Sazonalidade:

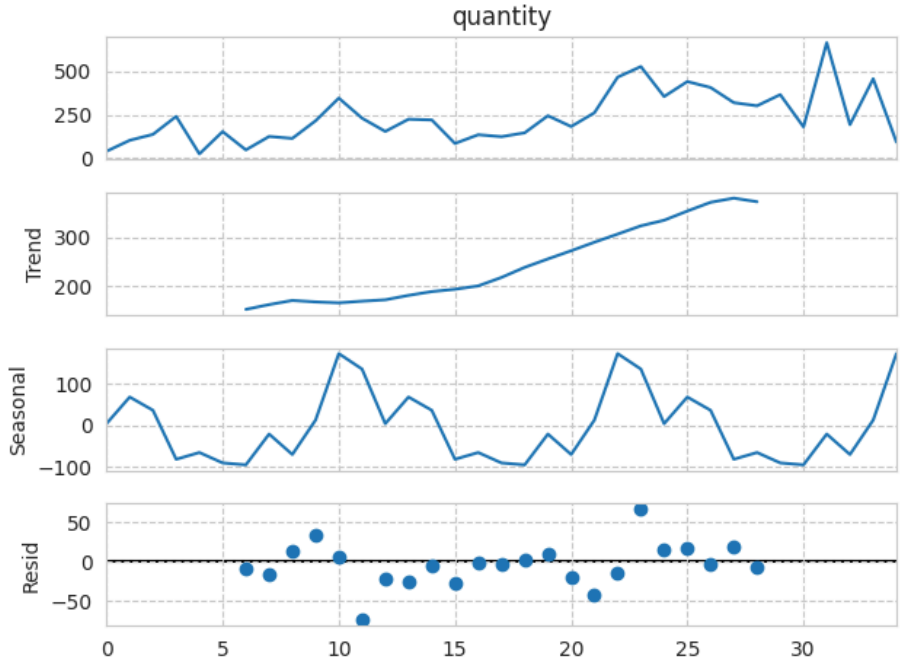
Durante a análise temporal das vendas, identificou-se a presença de sazonalidade nos dados gerais de produtos vendidos ao longo dos anos, com uma clara tendência de crescimento, conforme pode ser observado nos gráficos a seguir.



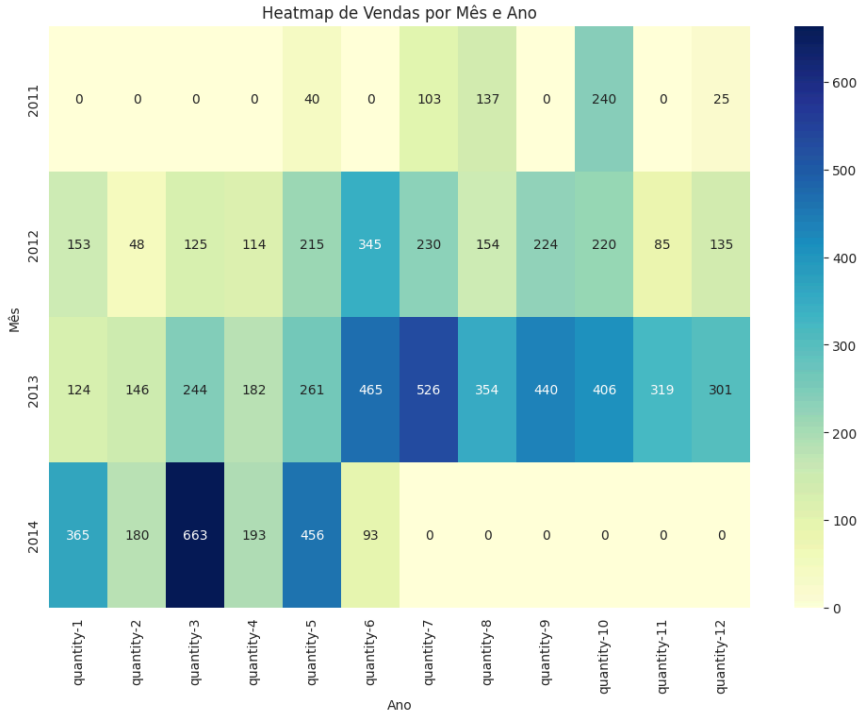
Entre os produtos vendidos pela empresa, aquele que apresentou maior demanda foi o **AWC Logo Cap**, com mais de 8.000 unidades vendidas no período de 2011 a 2014. Portanto, foi realizada uma análise de sazonalidade específica para esse produto.



A análise mostrou que também existe a presença de sazonalidade na venda desse produto, com uma significativa tendência de crescimento no final do 1º semestre e início do 2º, principalmente nos meses de maio a junho e de setembro a outubro. Esse padrão pode ser um indicativo do aumento de vendas em função das promoções de Black Friday, das comemorações de final de ano e, possivelmente, do período de férias escolares.



O mapa de calor a seguir mostra a intensidade na variação de produtos vendidos ao longo dos meses e anos.



Modelos Preditivos de Demanda:

Optou-se pelo uso de técnicas de **séries temporais hierárquicas** para a previsão de demanda em cada nível de agrupamento dos dados.

Séries temporais hierárquicas são abordagens que permitem modelar e prever dados em diferentes níveis de uma hierarquia. Essas técnicas são valiosas porque possibilitam análises detalhadas e previsões que consideram as relações entre diferentes níveis de granularidade, garantindo uma visão abrangente e precisa da demanda.

Elas ajudam a identificar padrões específicos em cada nível da hierarquia e a ajustar as previsões com base em informações agregadas e detalhadas, melhorando a precisão das previsões e otimizando a alocação de recursos.

Neste projeto, foram selecionadas os seguintes níveis hierárquicos:

1. Países (country)
2. Estados/Províncias (state_province)
3. Lojas (store_name)
4. Produtos (product_name)

Após a implementação da agregação dos dados por níveis hierárquicos, foram utilizados os algoritmos **AutoARIMA** e **Naive** para o treinamento dos modelos e respectivas previsões de demanda em cada nível hierárquico nos 3 meses seguintes da última data cadastrada no dataset.

O AutoARIMA foi escolhido por sua capacidade de identificar automaticamente os melhores parâmetros para modelos ARIMA, adaptando-se a padrões sazonais e tendências nos dados.

Já o Naive foi utilizado como uma abordagem simples e eficiente para estabelecer uma linha de base, com previsões baseadas na última observação disponível, oferecendo um comparativo direto com modelos mais complexos. Essas escolhas permitiram avaliar a precisão dos modelos e obter insights práticos para a previsão de demanda.

Os modelos foram treinados com as seguintes variáveis no horizonte de 3 meses:

- **unique_id:** ID único de cada nível hierárquico.
- **ds:** data (mês-ano) que ocorreu a venda do pedido.
- **y:** quantidade de produtos vendidos dentro desse nível hierárquico.

Para comparar os resultados dessa primeira modelagem, foram utilizadas as seguintes métricas:

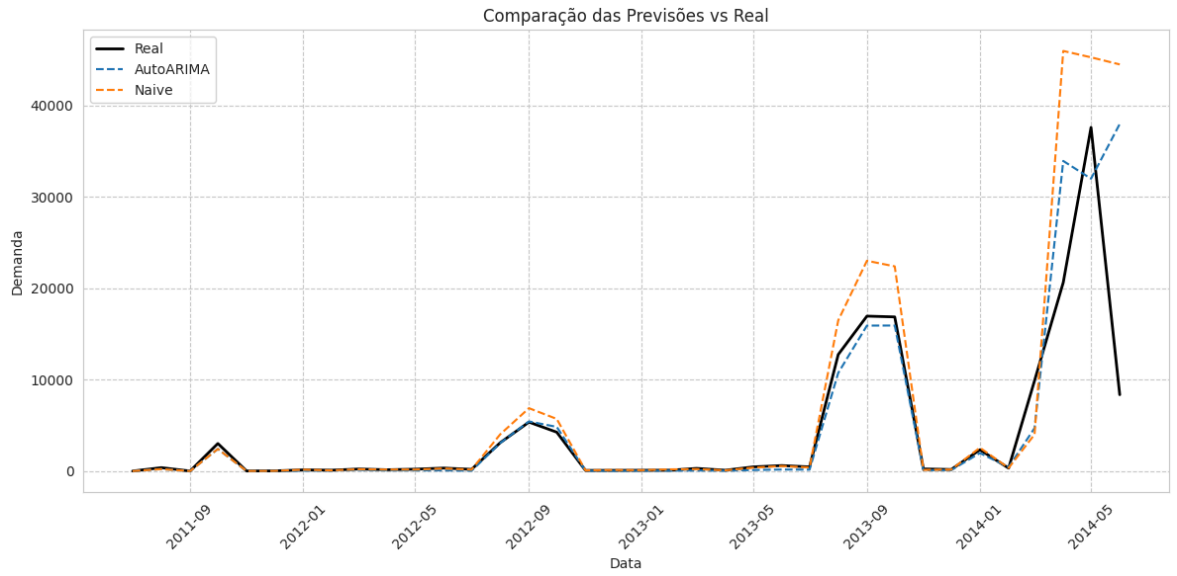
- **MAE (Erro Absoluto Médio):** mede o erro médio absoluto entre os valores previstos e os valores reais. Isto é, quanto em média os valores previstos estão afastados dos reais, sem considerar a direção do erro.

- **MSE (Erro Quadrático Médio):** mede o erro médio quadrático entre os valores previstos e os reais, dando maior peso a erros maiores. Penaliza mais os erros grandes, sendo útil para detectar grandes desvios.

Os resultados mostraram que o modelo AutoARIMA teve um desempenho melhor em termos de MAE em comparação ao modelo Naive, o mesmo vale em termos de MSE. Isso sugere que o AutoARIMA fornece previsões mais precisas para a demanda do que o algoritmo Naive.

A tabela a seguir mostra as métricas de desempenho obtidas para cada modelo e o gráfico mostra a comparação entre os valores reais e previstos por cada um dos modelos:

Modelo	MAE		MSE	
	Soma	Média	Soma	Média
AutoARIMA	65768.68	6.84	73375.32	7.64
Naive	70264.67	7.31	77071.43	8.02



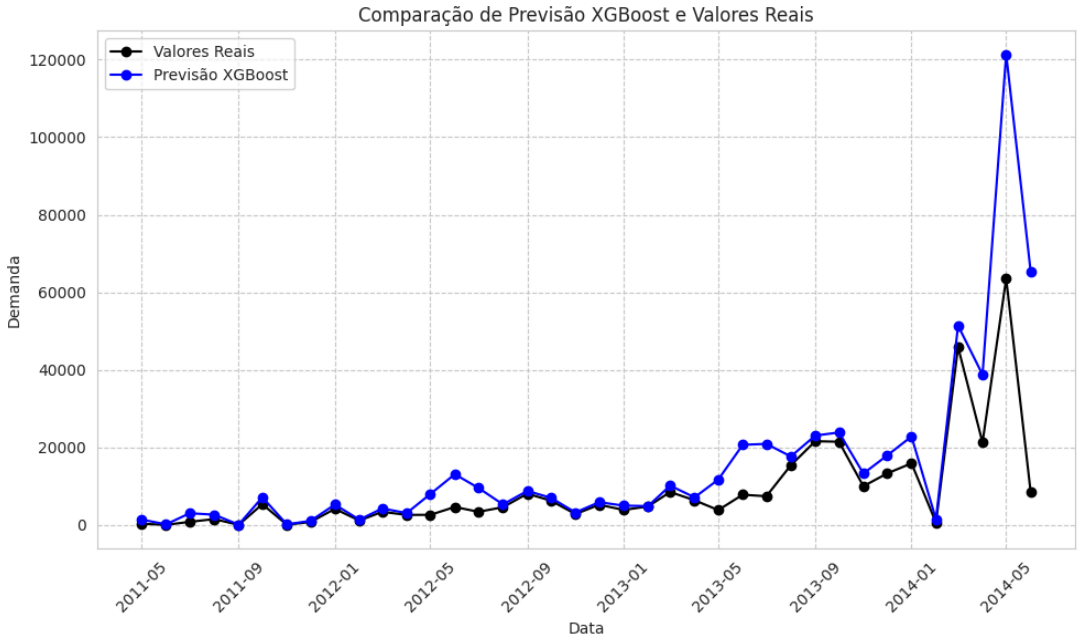
Uma outra alternativa para lidar com previsão de demanda em séries hierárquicas é através de modelos de regressão, que permitem incorporar múltiplas variáveis explicativas no processo de previsão, como características de produtos, regiões e comportamentos sazonais.

Ao utilizar modelos de regressão é possível capturar relações complexas e interações entre variáveis que podem não ser facilmente identificadas por métodos

tradicionais de séries temporais. Além disso, essa abordagem oferece maior flexibilidade para ajustar o modelo de acordo com mudanças nas condições de mercado e novos padrões de dados.

Os benefícios incluem a capacidade de prever em diferentes níveis de granularidade sem precisar agregar ou desagregar manualmente, maior robustez na incorporação de variáveis categóricas e numéricas e a facilidade de ajustar o modelo para diferentes cenários de previsão, como a introdução de novos produtos ou mudanças sazonais.

Portanto, para validar o ganho de performance na implementação de um modelo de regressão, foi selecionado o algoritmo **XGBoost Regressor** para prever a demanda de produtos nos diferentes níveis hierárquicos. Para visualizar melhor esses resultados, o gráfico abaixo faz uma comparação entre os valores reais e os valores previstos pelo XGBoost.



Na tabela a seguir foram agrupadas e comparadas as métricas de desempenho de cada um dos modelos implementados neste projeto a fim de identificar aquele que apresentou melhor performance e que representa uma melhor solução para previsão de demanda dos produtos da Adventure Works.

Para essa análise, foram incluídas também as seguintes métricas:

- **MAPE (Erro Percentual Absoluto Médio):** mede o erro percentual médio entre os valores previstos e reais. Expressa o erro em termos percentuais, útil para comparar previsões em diferentes escalas.
- **RMSE (Raiz do Erro Quadrático Médio):** é a raiz quadrada do MSE, trazendo o erro de volta à mesma escala dos dados. Mede a magnitude média dos erros, sendo sensível a grandes desvios, mas em uma escala mais interpretável.

Modelo	MAPE	MAE	MSE	RMSE
AutoARIMA	0.52	1728.7	3.11×10^7	5575.85
Naive	0.37	2665.25	5.91×10^7	7684.35
XGBRegressor	1.82	5.54	3547.1	59.56

É possível observar que o XGBoost Regressor apresentou quase todas as métricas abaixo dos demais modelos, com exceção do MAPE. O MAPE é muito sensível quando os valores reais da demanda são baixos. Isso ocorre porque o MAPE calcula o erro percentual, e se o valor real de y for pequeno ou próximo de zero, o erro percentual pode se tornar muito grande, mesmo que o erro absoluto (MAE) seja relativamente pequeno.

O XGBoost pode estar superestimando ou subestimando pequenos valores, o que leva a um aumento desproporcional no MAPE. Além disso, o XGBoost é um modelo de machine learning baseado em árvores, o que significa que ele pode capturar padrões complexos nos dados, especialmente não lineares, o que tende a melhorar o MAE, MSE e RMSE.

No entanto, isso não necessariamente melhora o MAPE se houver variações bruscas em determinadas faixas de valores. AutoARIMA e Naive são modelos de séries temporais mais simples, que podem gerar previsões mais estáveis, resultando em um MAPE menor, mas erros absolutos maiores. Por isso o comportamento identificado nas métricas apresentadas anteriormente.

Demanda por Centro de Distribuição:

Após uma reformulação dos seus processos, a Adventure Works passou a dividir os seus novos centros de distribuição de forma diferenciada nos Estados Unidos, onde estes foram segmentados por províncias, enquanto nas demais localidades a segmentação se manteve em países. Nesse sentido, é importante entender como a demanda de produtos se comporta dentro desses agrupamentos.

Pensando nisso, os dados previstos foram agrupados nesses dois grupos e em seguida foi calculado o percentual de crescimento de demanda para cada um deles, gerando os seguintes resultados:

Agrupamento	Demanda Anterior (3 meses antes)	Demanda Prevista (próximos 3 meses)	Aumento de Demanda (%)
Outros Países	26.642	104.491	292.2
Províncias (EUA)	33.600	66.546	98.05

Portanto, os centros de distribuição dos EUA apresentaram um crescimento em demanda de 98%, enquanto os centros dos demais países apresentaram um crescimento de demanda de 292%, mais que o dobro do grupo anterior.

Estimativa de Demanda de Zíperes:

Considerando que um novo fornecedor de luvas assumiu toda a produção global, foi necessário estimar a quantidade de zíperes que deverá ser solicitada para os próximos 3 meses. Para isso, foi realizada uma análise detalhada da demanda, levando em conta as previsões de vendas e o consumo de 2 zíperes em cada unidade produzida.

Para isso, foi necessário filtrar os dados previstos somente para os produtos da subcategoria “Gloves”, que engloba os produtos Half-Finger Gloves (S, M, L) e Full-Finger Gloves (S, M, L). Nos 3 meses previstos, foi identificada a demanda de **1.806 pares de luvas**, gerando uma estimativa de **3.612 zíperes** para que sejam fabricadas. Dessa forma, é necessário que o fornecedor realize a compra dessa quantidade de zíperes, considerando uma margem de erro por segurança. Esses resultados foram obtidos a partir dos valores previstos no modelo XGBoost.

6. Riscos

Existem alguns riscos associados ao projeto que devem ser gerenciados de forma proativa. Isso inclui a resistência de alguns stakeholders ao projeto, dificuldades no acesso aos dados e a aceitação dos modelos preditivos. A mitigação desses riscos será crítica para o sucesso do projeto.

A seguir serão listados alguns riscos identificados na implementação do projeto:

I. **Qualidade dos dados:** Idealmente é necessário que a qualidade dos dados de entrada sejam consistentes e precisos para garantir bons resultados. No entanto, o conjunto de dados da Adventure Works apresentou pouca diversidade e volume de dados quando analisado os agrupamentos hierárquicos para criação do modelo preditivo, o que implicou em um baixo desempenho dos modelos selecionados.

7. Entregáveis

1. **Modelo Conceitual do Data Warehouse**
2. **Dashboard de Vendas**
3. **Relatórios de Análise Exploratória e Temporal dos Dados**
4. **Modelos de Predição de Demanda**
5. **Relatório Detalhado do Projeto**

8. Próximos Passos

1. **Alinhamento Final de Escopo**: Reunião com as principais partes interessadas para definir o cronograma e garantir o entendimento claro do valor estratégico do projeto.
2. **Validação e Ajustes**: Teste do modelo preditivo com as equipes envolvidas e ajustes conforme necessário.