Modelação de um Data Warehouse

Grupo 17

Afonso Gama

55857

Eduardo Carneiro

62515

Guilherme Rosário

62543

Marco Viana

62550

Integração e Processamento Analítico de Informação

Mestrado em Engenharia Informática

Índice

Conteúdo

1. Introdução

No presente projeto, proposto na unidade curricular de Integração e Processamento Analítico de Informação da Faculdade de Ciências da Universidade de Lisboa, pretende-se explorar a utilização de diversas ferramentas com o objetivo de modelar e construir um Data Warehouse. O modelo a apresentar tem como objetivo compreender e otimizar os processos de negócio envolvidos na operação de uma pizzaria, desde o registo dos pedidos até à análise de padrões de consumo.

Na primeira etapa, descrita no relatório que se apresenta, o foco principal passa por identificar as fontes de dados relevantes para o processo de negócio em questão, analisando os valores e erros nos campos de dados de cada fonte, assim como o seu tratamento. Disponibiliza-se um diagrama que tem como propósito representar as ligações estabelecidas correspondentes aos dados pós-processamento. Por fim define-se um processo de negócio para tirar partido dos dados recolhidos e estabelecem-se três questões analíticas fundamentais como proposta de futura análise.

Para a análise dos dados foi utilizada a linguagem de programação *Python*, através da qual foram gerados plots e histogramas, os quais forneceram insights acessíveis acerca da informação presente no dataset. Por fim, a representação das informações num diagrama permitiu uma fácil visualização da ligação entre as diferentes fontes de dados e a identificação das subsequentes hierarquias.

Na segunda etapa, a qual se inicia na secção 7 do presente relatório, efetuou-se a modelação dimensional. Primeiramente, apresenta-se a tabela de factos e o respetivo grão, assim como as características inerentes à mesma, passando seguidamente para uma análise mais detalhada das dimensões associadas à tabela. Apresenta-se, no fim, um diagrama em estrela que permite, de forma visual e clara, entender as ligações estabelecidas e pormenores destas "relações".

2. Análise do Dataset

O conjunto de dados escolhido retrata o agrupamento de operações da Pizzaria Plato´s, uma pizzaria com inspirações gregas localizada em New Jersey, que conta com um total de 15 mesas e 60 cadeiras, e tem como objetivo continuar a evoluir. Como tal, ao longo do último ano, têm vindo a recolher vários tipos de dados que se demonstram em seguida.

Sendo completamente público, o dataset original está disponível no website Maven Analytics, que é uma plataforma online onde analistas e investigadores podem aprender novos conhecimentos na área de análise de dados, oferecendo ainda a possibilidade de contacto entre membros. Na plataforma é possível aceder ao ”Data Playground” onde, para além do que serve de base para a realização deste relatório, se armazenam outros conjuntos de dados apropriados para análise.

Na análise a efetuar é utilizada uma versão obtida através da plataforma Kaggle, que possui um tamanho total de 5,33MB, no qual todas as instâncias se apresentam na mesma tabela de forma a tornar mais fácil o processamento dos dados por parte do grupo. Com um número ordem as 48600 linhas e com um total de 12 colunas, disponibilizam-se dados como a data e tempo de uma venda, as pizzas a servir com vários detalhes como tipo ou quantidade entre outros. Na seguinte secção encontra-se uma tabela na qual é efetuado o escrutínio de todas as colunas.

2.1 Tipos de Dados e Explicação das Colunas

Tabela 1: Tipos de Dados das Colunas e respetiva Explicação

Colunas	**Tipos de Dados**	**Explicação**
Order ID	Integer	Identificador de cada pedido efetuado
Order Date	Date	Data da encomenda
Order Time	Hour	Horas da encomenda
Total Price	Float	Preço total da encomenda, redundância equivalente a (Preço unitário * Quantidade)
Order Details Id	Integer	Identificador dos detalhes de cada pedido efetuado
Pizza ID	String	Identificador de cada pizza colocada na encomenda
Quantity	Integer	Quantidade de cada pizza encomendada
Unit Price	Float	Preço da encomenda, em Dólar
Pizza Size	String	Categoria para o tamanho da pizza
Pizza Category	String	Identificador do tipo de pizza
Pizza Ingredients	List of String	Ingredientes da pizza encomendada
Pizza Name	String	Nome da pizza encomendada

2.2 Visualização dos Dados

2.2.1 Estatísticas individuais de cada coluna

Após ser analisado o tipo de dados e o significado de cada coluna, foram analisadas as estatísticas individuais das mesmas, como forma a compreender mais detalhadamente as variáveis em estudo. Para tal, foram feitas duas tabelas, na tabela 2 estão representadas as variáveis não contínuas (variáveis que podem assumir apenas valores específicos e distintos), e na tabela 3 estão representadas as variáveis contínuas (variáveis que podem assumir uma infinidade de valores dentro de um intervalo específico)

Tabela 2: Variáveis Não-Contínuas

Variável	**Número Valores Únicos**	**Valor Mais Frequente**
pizza id	91	big meat s
pizza category	4	Classic
pizza name	32	The Classic Deluxe Pizza
pizza size	5	L
order date	358	2015-11-26
order time	16382	12:32:00

Tabela 3: Variáveis Contínuas

Variável	Número de Entradas	Min	Max	Média	Std
order details id	48620	1	48620	nan	nan
order id	48620	1	21350	nan	nan
quantity	48620	1	4	1.02	0.14
unit price	48620	9.75	35.95	16.49	3.62
total price	48620	9.75	83	16.82	4.44

2.2.2 Distribuição de Valores por Coluna

De modo a obter uma perspetiva geral das distribuições dos valores para cada coluna, produzimos histogramas com as mesmas. Nestas figuras, é possível observar a contagem de quantas vezes cada valor aparece (nas variáveis contínuas foi efetuada automaticamente a operação de criação de *bins*), bem como o KDE (*Kernel Density Estimation*), i.e., a curva com uma estimativa da função de densidade probabilística. Esta curta visualização permite rapidamente perceber algumas tendências nos dados, sem necessitar de quaisquer análises muito complexas.

Através da visualização das figuras apresentadas é possível retirar algumas conclusões de alto nível acerca do dataset.

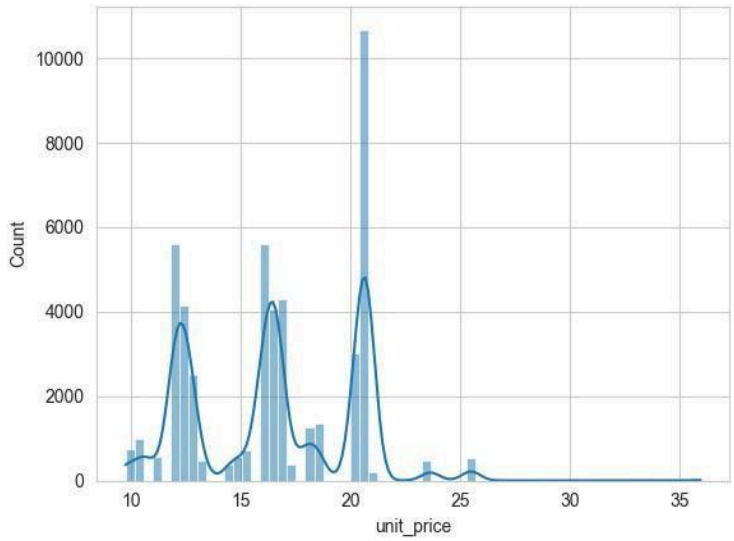

Figura 2: Histograma dos Preços Unitários das Pizzas

- Em relação ao preço unitário de cada pizza (2), é percetível a existência de 3 bandas de preços, isto é, 3 intervalos com maior incidência dos clientes: 12 a 14 euros, com maior ênfase em valores mais perto dos 12 euros; 17 a 18 euros, com maior destaque à volta dos 17 euros; e, finalmente, 20 a 22 euros, com um realce claro em torno dos 20 euros.

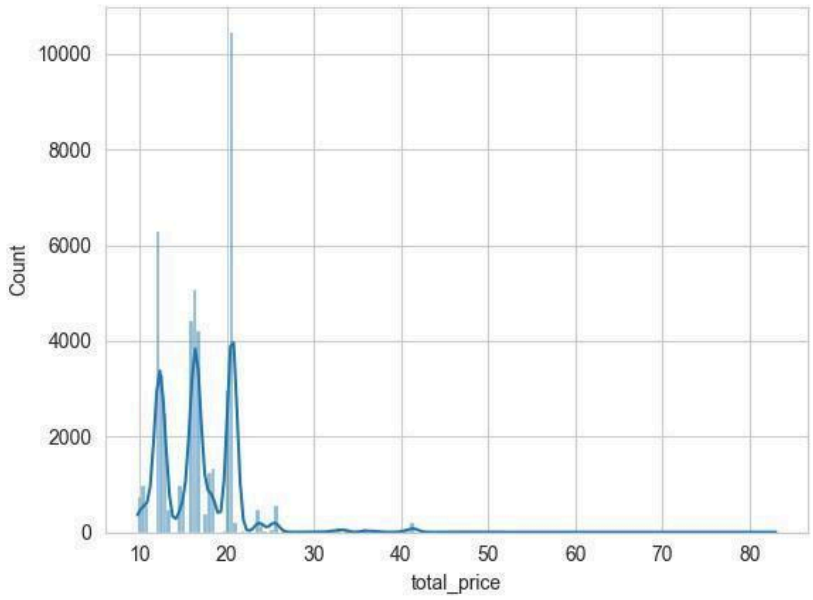

Figura 3: Histograma do Valores Totais dos Pedidos

- Na Figura 3, representativa dos valores totais referentes a cada encomenda, é possível observar as mesmas tendências explicadas para a Figura 2, isto é, 3 bandas de preços. Estas bandas podem refletir dois cenários: os pedidos com tamanhos de pizzas diferentes, o que altera o preço total do pedido, ou 3 tipos diferentes de clientes que frequentam o restaurante. De qualquer maneira, é necessária uma análise mais profunda para consistentemente fazer estas afirmações.

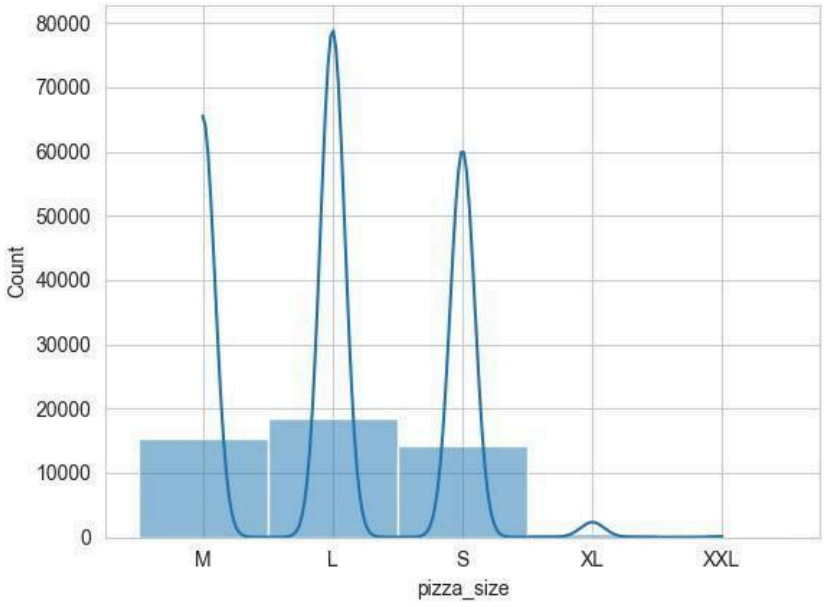

Figura 4: Histograma dos Tamanhos de Pizza

- No histograma da Figura 4 compreende-se que há uma clara incidência dos clientes para 3 tipos de tamanhos de pizza: *M, L* e *S*. Tendo em conta as análises realizadas aos histogramas anteriores, é agora mais credível que as ”bandas” de preços mencionadas anteriormente se refiram, na realidade, a encomendas com tamanhos diferentes de pizza, sendo estes os 3 mais comuns. No entanto, e como já foi referido, seria necessária uma análise mais complexa para extrair conclusões mais fundamentadas.

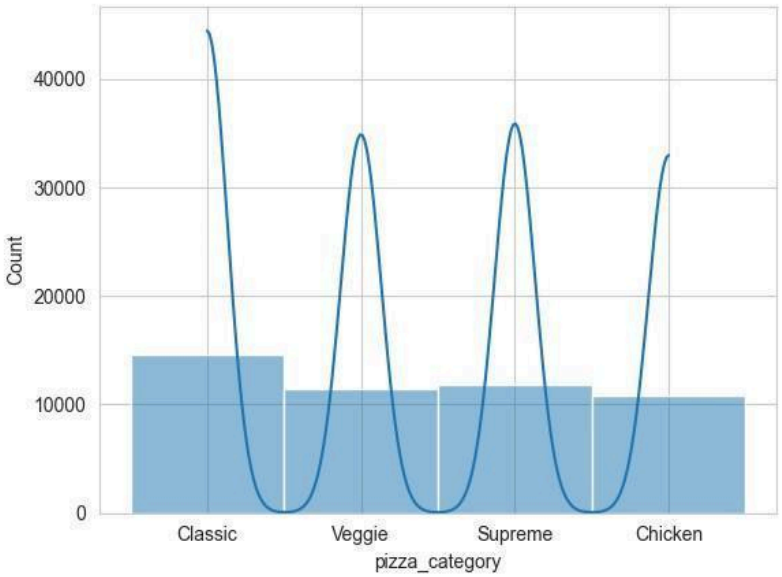

Figura 5: Histograma das Categorias de Pizza

- A respeito das categorias de pizzas, na Figura 5, temos o que seriam valores expectáveis: a pizza do tipo *Classic* com maior procura, e os restantes tipos bastante equilibrados. Todavia, a discrepância entre a categoria mais pedida e as restantes é, possivelmente, negligenciável.

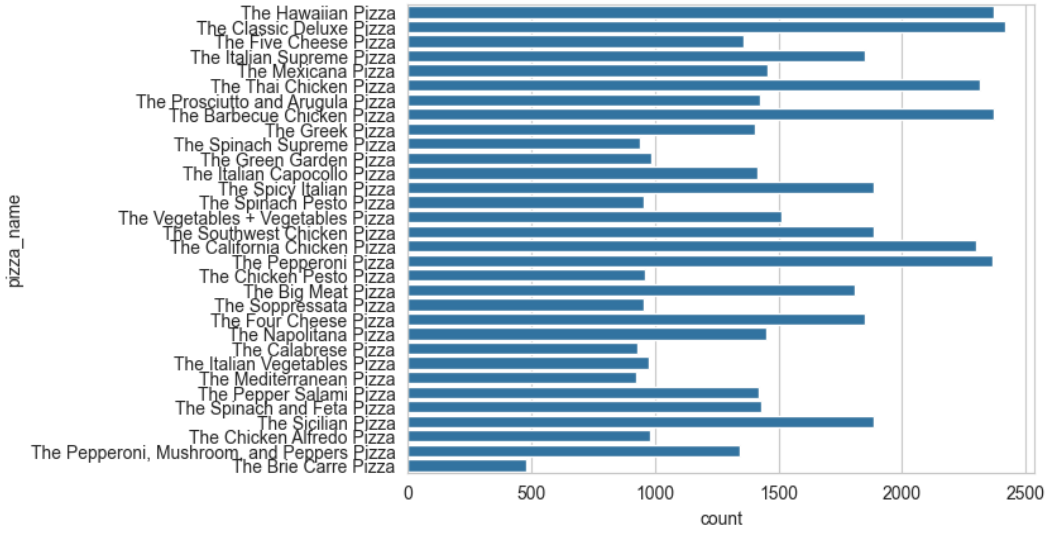

Figura 6: Histograma dos Nomes de Pizza

- A Figura 6 demonstra a frequência com que cada pizza foi pedida. É possível observar que

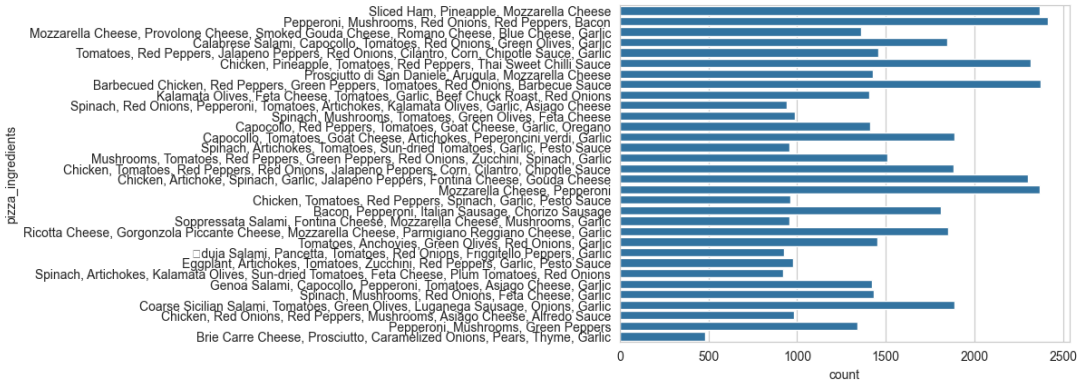

Figura 7: Histograma dos Conjuntos de Ingredientes

- Na Figura 7 pode ver-se os conjuntos de ingredientes mais pedidos. Como era de esperar, coincide com as pizzas mais pedidas.

2.2.3 Correlação entre Variáveis

A seguinte figura apresenta as correlações entre as variáveis contínuas do dataset.

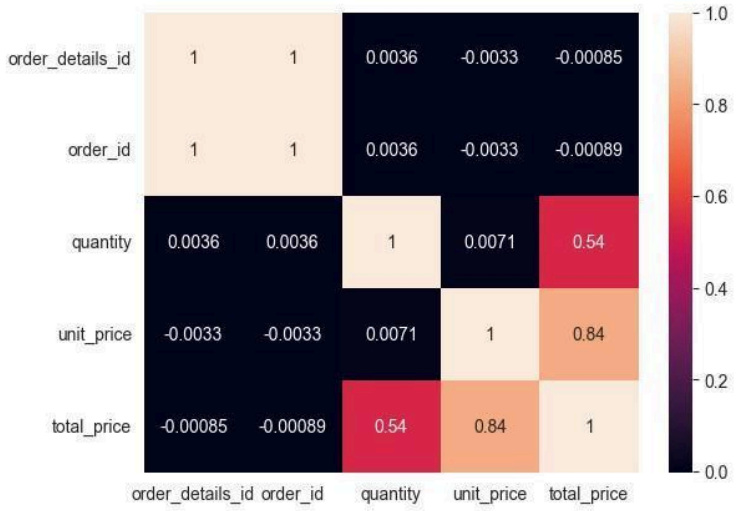

Figura 7: Matriz de Correlação entre Variáveis

É percetível que, por exemplo, a variável *unit_price* tem uma correlação de 0,84 com a variável *total_price*. Este valor é expectável, tendo em conta que muito provavelmente muitas encomendas apenas têm uma pizza, o que leva a que estas variáveis sejam iguais recorrentemente. Esta razão também explica a correlação de 0.54 entre *quantity* e *total_price.* Relativamente às variáveis dos ID's apresentam uma correlação de 1, mas esta correlação é irrelevante tendo em conta que estas variáveis não representam nenhum atributo acerca dos dados.

3. Uniformização e Limpeza dos Dados

O dataset contém uma particularidade em relação aos datasets comuns encontrados no Kaggle, uma vez que não existem valores em falta (nulls), nem valores duplicados, o que facilitou bastante a análise inicial do dataset. Quanto aos nomes das colunas, estão bem formados e são autoexplicativos, pelo que não há necessidade de proceder a qualquer alteração.

No que diz respeito aos valores atípicos, trata-se de um tema sensível. Dada a natureza das variáveis, a deteção de valores atípicos pode ser complexa ou mesmo impossível. Por exemplo, para a variável *Preço unitário*, é difícil decidir se um valor é ou não um valor atípico apenas com base em estatísticos como os quantis ou métodos de agrupamento, dado que uma disparidade de valores no preço não significa que um deles esteja errado. Dito isto, decidimos não fazer qualquer remoção de valores atípicos. No entanto, efetuamos alterações em algumas colunas para facilitar a análise futura e a manipulação de objetos: divisão dos timestamps de data/hora na sua própria coluna identificada, ou seja, a hora tem a sua própria coluna, o dia tem a sua própria coluna, etc…

4. Diagrama

O diagrama que se apresenta permite clarificar as relações adequadas ao conjunto de dados escolhido. A tabela ”Pedidos” propõe armazenar um registo de todos os pedidos efetuados no restaurante, registando a data e hora dos mesmos e também o valor total gasto. Já na tabela ”Detalhes do Pedido” encontram- se as pizzas e a quantidade das mesmas de um determinado pedido. Pizzas estas que se registam na tabela ”Pizzas”, onde a categoria, tamanho e preço destas é atribuído. Por fim, a tabela ”Tipos de Pizza” disponibiliza os ingredientes a utilizar numa determinada pizza assim como o seu nome, sendo esta inserida numa categoria estabelecida pelo restaurante.

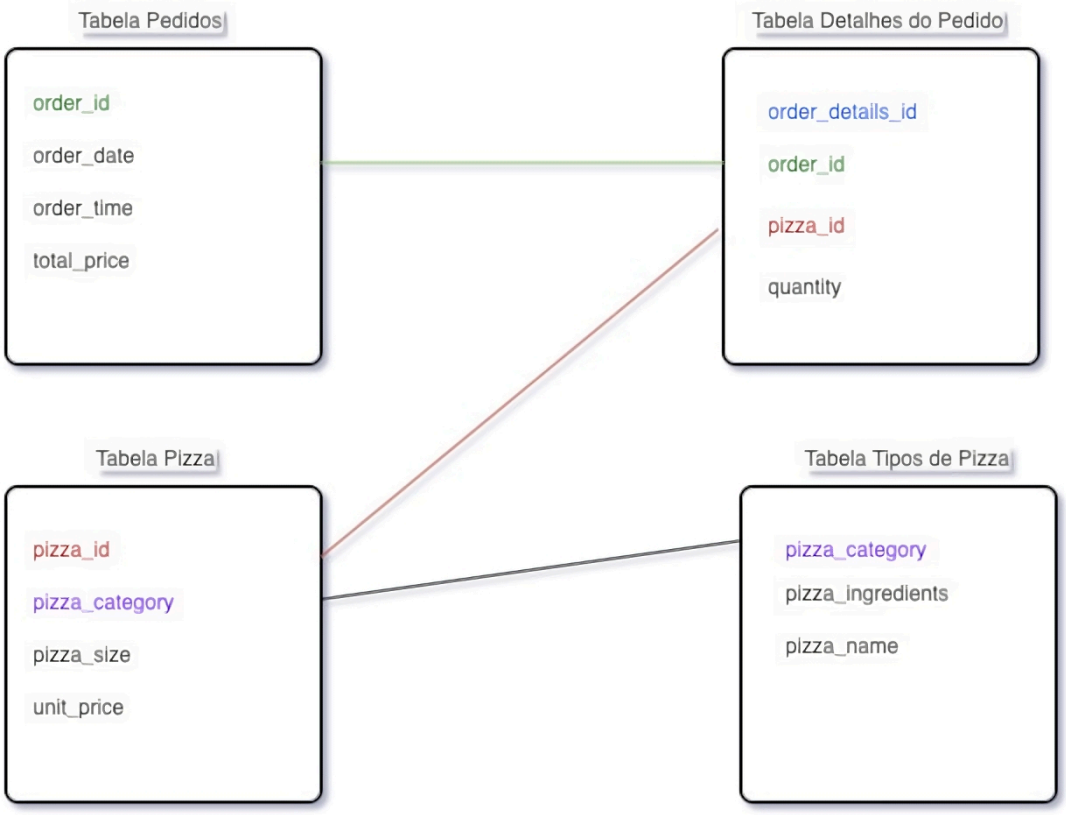

Figura 8: Diagrama representativo do conjunto de dados

5. Processo de Negócio

É crucial, em qualquer negócio, acompanhar, analisar e compreender as vendas realizadas, visando explorar potenciais incrementos de rendimentos e facilitar a identificação de falhas. Nesse sentido, foi selecionado um processo de negócio centrado na performance de vendas, o qual pretende apoiar análises futuras, de modo a fornecer análises fundamentais para determinar o rumo do negócio, investimentos e/ou a resolução de eventuais problemas.

Garantir a qualidade dos dados é fundamental, uma vez que informações imprecisas ou incompletas podem resultar em decisões erradas. Para maximizar as vendas de pizzas, é essencial otimizar os processos de negócio, garantindo eficiência, qualidade e satisfação do cliente.

Primeiramente, é necessário realizar uma pesquisa de mercado abrangente para compreender o público-alvo, a concorrência local, a procura do mercado e identificar os dados demográficos, preferências e comportamentos de compra do cliente-alvo. Além disso, é crucial desenvolver um menu diversificado e atrativo, introduzindo pizzas únicas que diferenciam a marca da concorrência. A qualidade dos ingredientes também é crucial, utilizando produtos frescos, carnes de primeira qualidade e queijos autênticos para garantir o sabor das pizzas. Medidas de controle de qualidade devem ser implementadas para manter a consistência e frescura dos ingredientes. O tempo é um fator fundamental, portanto, é importante conceber um processo de produção eficiente para minimizar os tempos de espera e garantir um serviço rápido durante as horas de maior movimento, mantendo a qualidade dos produtos. A equipa de cozinha deve ser treinada para seguir receitas e procedimentos padronizados, garantindo eficiência sem comprometer a qualidade. Um sistema de pedidos simples e acessível é essencial. Várias opções de entrega devem ser disponibilizadas, incluindo entregas internas e parcerias com serviços de entrega de terceiros, para maximizar as vendas. É fundamental oferecer um serviço excecional ao cliente em todos os pontos de contacto, com uma equipa simpática e bem informada, cumprimento imediato dos pedidos e capacidade de resposta às perguntas e comentários dos clientes. Um ambiente acolhedor e confortável deve ser garantido para os clientes que jantam no restaurante. Uma estratégia de marketing abrangente é essencial para promover a

pizzaria através de vários canais, incluindo redes sociais, publicidade local, marketing por e-mail e parcerias com outras empresas. Promoções, descontos e programas de fidelidade devem ser oferecidos para incentivar a fidelização e atrair novos clientes.

É também importante solicitar feedback dos clientes de forma fácil para identificar áreas de melhoria e utilizar esse feedback para tomar decisões e aprimorar continuamente os processos comerciais, menus e práticas de atendimento ao cliente. Com o dataset que está a ser utilizado, que nos fornece todo o tipo de informações como, as pizzas que estão disponíveis, as mais pedidas, as menos pedidas, os seus valores, etc, facilita a tarefa de analisar todos estes pontos para se poder maximizar as vendas e a qualidade dos produtos, de forma a manter clientes e a atrair novos.

6. Questões Analíticas

1. Como é que a composição de ingredientes de cada tipo de pizza influencia a sua popularidade e o valor médio de vendas?

Esta é, possivelmente, a questão mais importante de todas. Permite perceber quais são os tipos de pizzas que os clientes mais procuram e os que menos procuram, o que auxilia, por exemplo, a eventualmente remover algumas pizzas do menu. Outro ponto fundamental desta questão é que possibilita, implicitamente, a análise dos ingredientes mais utilizados na cozinha. Deste modo, é possível entender quais são os ingredientes que necessitam de mais stock, o que é uma questão primordial num restaurante, principalmente quando estão na equação ingredientes com prazos de validade curtos, os quais têm de ser pedidos diariamente ou semanalmente. Também, através da análise dos ingredientes mais usados, é proporcionada a possibilidade de orquestrar novas pizzas agregando os ingredientes mais apreciados pelos clientes.

2. Quais são os dias e horas mais movimentados do restaurante?

Esta análise propõe explorar a influência do espaço temporal na movimentação do restaurante, sendo possível apresentar questões pertinentes ao nível dos dias da semana como, quais os dias com menor movimento, o que poderá posteriormente ser utilizado como fundamento para oferecer descontos em dias mais calmos, bem como reduzir a quantidade de ingredientes encomendados para esses dias. Assim como ao nível de horários, onde é de interesse entender, as horas com maior número de pedidos e/ou as horas com maior quantidade de pizzas encomendadas. Este diagnóstico possibilita a melhor gestão de recursos humanos no restaurante, bem como a possibilidade de fazer descontos em horas menos frequentadas de modo a incentivar a procura.

3. De que maneira a pizzaria está a utilizar a sua capacidade de lotação?

Será com certeza pertinente para qualquer restaurante perceber de que maneira é utilizado pelos clientes o espaço de refeições e esta pizzaria não é exceção. De que maneira podemos organizar as mesas e cadeiras disponíveis, são estas suficientes ou demais para os pedidos efetuados, recebem mais clientes a comer sozinhos ou acompanhados por outras pessoas,

existe todo um conjunto de informações possíveis de retirar neste tópico, que ao podem ser determinantes para a otimização do processo deste e de qualquer outro restaurante.

7. Tabela de factos e grão

A tabela de factos é do tipo transacional, ou seja, cada entrada na tabela representa uma transação efetuada. Cada transação corresponde a uma pizza pedida e varia em termos de quantidade. Um pedido é caracterizado por um conjunto de transações, isto é, um pedido tem associada uma ou mais transações, nas quais cada uma delas corresponde a pizzas diferentes pedidas.

Apresenta-se, como exemplo para elucidar esta explicação, o caso em que um cliente pediu duas pizzas *pepperoni* e uma *four cheese*. Apesar de ser o mesmo pedido, identificado pelo campo *Order Key*, presente na fact table como chave estrangeira da dimensão *Order*, são duas transações diferentes na tabela de factos, uma para as duas pizzas de *pepperoni* e uma para a de *four cheese.*

A tabela de factos, denominada como *factOrderDetails* inclui um total de três dimensões, sendo estas: *Pizza, Order* e *Date*, descritas em mais pormenor na secção 8. Estão presentes na tabela também duas medidas*, Quantity* – que representa a quantidade de pizzas de uma transação - *e Total Price Transaction* – que representa o preço total de uma transação – estas são de extrema importância para dar resposta as questões analíticas determinadas na etapa anterior e úteis, no ponto de vista do grupo, a possíveis novos pontos de interesse a estudar/analisar.

A tabela *factOrderDetails* possui um 48620 linhas.

Tabela 4: Descrição dos campos de dados: Parte 1 - Identificadores e atributos

Campo	Descrição dos dados	Origem dos dados
Transaction	Chave Primária (identificador do pedido)	ID gerado sequencialmente
Pizza Key	Chave Estrangeira (Dimensão Pizza)	–
Data Key	Chave Estrangeira (Dimensão Date)	–

Order Key	Chave Estrangeira (Dimensão Order)	–

Tabela 5: Descrição dos campos de dados: Parte 2 - Medidas aditivas

Campo	**Descrição**	**Origem dos dados**	**Valores**
Quantity	Quantidade de pizzas presentes na transação	Tabela Detalhes do Pedido	1 - 4 Média: 1.02
Total Price Transaction	Preço total da transação	Tabelas Pedidos	9.75 - 83 Média: 16.82

8. Dimensões

A criação das dimensões são um passo essencial na modelação, sendo esta mesma fulcral para especificar o detalhe das informações relativas a cada transação (tendo em conta que é isso que define o nível de detalhe refletido na tabela de factos). Como referido anteriormente, foram criadas quatro dimensões com características diferentes que permitem caracterizar cada linha da tabela de factos de forma detalhada. Em seguida, são pormenorizadas todas as dimensões, de modo a elucidar o seu papel e as suas características no data warehouse.

8.1 Dimensão data – DimDate

A dimensão data é crucial em data warehouses e, no contexto deste projeto, é utilizada para detalhar as vendas de pizzas, com vista a capturar todos os aspetos temporais relevantes de cada transação. Esta dimensão é estruturada com uma Chave Primária, a *Data Key*, que vincula cada venda registada na tabela de factos a uma ocorrência temporal.

Os atributos desta dimensão foram elaborados utilizando-se as informações de data e hora das vendas. Os dados iniciais - “Day”, “Month”, “Year”, “Hour”, “Minute”, “Second” - que serviram de base para a criação de atributos adicionais que aumentam a utilidade do modelo e proporcionam maior flexibilidade analítica. Deste modo, foram adicionados o “Day of Week Number” (dia da semana em forma numérica), “Day Name Of Week” (Nome do dia da semana) e “Weekend Indicator” (indicar se é ou não fim de semana). Estes detalhes são importantes para diferentes tipos de análises.

Os aprimoramentos na dimensão data não só fortalecem a capacidade analítica do data warehouse, como também oferecem análises estratégicas que podem ser traduzidos em ações concretas, de forma a impulsionar o crescimento e a adaptação do negócio num mercado competitivo.

Hierarquias: Year > Month > Day > Minute > Second

Day Number Of Week > Day Name of Week

Tamanho: 16 382 linhas

Tabela 6: Dimensão Data

Campo	Descrição dos dados	Origem dos dados	Valores
Data Key	Chave Primária	ID gerado sequencialmente	1 - x
Day	Número do dia do pedido	Tabela Pedidos	1 - 31
Month	Número do mês do pedido	Tabela Pedidos	1 - 12
Year	Número do ano do pedido	Tabela Pedidos	2015
Hour	Hora do pedido	Tabela Pedidos	0 - 24
Minute	Minuto do pedido	Tabela Pedidos	0 - 59
Second	Segundo do pedido	Tabela Pedidos	0 - 59
Day Number Of Week	Dia da semana em forma numérica (sendo 1 segunda-feira e 7 domingo)	Tabela Pedidos	1 - 7
Day Name Of Week	Nome do dia da semana	Tabela Pedidos	Monday - Sunday
Weekend Indicator	Indicador se é ou não fim de semana	Tabela Pedidos	No ou Yes

8.2 Dimensão Pizza – DimPizza

A Dimensão Pizza é fundamental para entender as características específicas de cada produto vendido neste restaurante. Esta dimensão facilita a análise detalhada das preferências do consumidor, relacionando o impacto do tamanho das pizzas, ingredientes e preço no desempenho de vendas.

Sendo também uma boa ferramenta para a análise de desempenho de produtos, permitindo que a gestão do restaurante ajuste ofertas baseadas em dados concretos sobre preferências de clientes e performance de vendas de diferentes tipos e tamanhos de pizzas.

Esta dimensão é composta pelos seguintes atributos: “Pizza Category”, “Pizza Name”, “Pizza Size”, “Unit Price” e “Ingredientes”.

Hierarquia: Pizza Category > Pizza Name > Pizza Size > Unit Price

Tamanho: 91 linhas

Tabela 7: Dimensão Pizza

Campo	Descrição dos dados	Origem dos dados	Valores
Pizza Key	Chave Primária	ID gerado sequencialmente	1 - x
Pizza Category	Chave Estrangeira (Identificador do tipo de categoria onde a pizza se insere)		1- 4
Pizza Name	Nome da pizza		Exemplo: “The Italian Supreme Pizza”
Pizza Size	Tamanho da pizza		“S”;“M”;”L”;”XL”;XXL”

Unit Price	Preço unitário por pizza		9.75 - 35.95
Ingredientes	Ingredientes que farão parte de uma determinada pizza		“Tomatoes”; “Garlic”; “Beef Chuck Roast”;

8.3 Dimensão Categoria – DimCategory

A dimensão Categoria, é caracterizada por ser um *outrigger*. Esta, organiza as pizzas em grupos baseados na sua categoria, permitindo, por exemplo, análises detalhadas sobre quais as categorias de pizzas mais comuns entre os consumidores. Esta dimensão é útil para a composição e organização da dimensão Pizza.

Esta dimensão apresenta apenas um atributo, sendo este o “Pizza Category”, sendo o tipo de categoria onde a pizza se insere.

Essa dimensão ajuda na análise de tendências de consumo, identificando quais categorias são mais populares. Por exemplo, é possível explorar se pizzas vegetarianas vendem mais durante certos meses do ano ou eventos específicos. Além disso, a dimensão Categoria permite a comparação de desempenho de vendas entre diferentes tipos de pizzas, fornecendo dados valiosos para decisões relacionadas a promoções, introdução de novos produtos ou ajustes no menu existente.

Hierarquia: Category Key > Pizza Category

Tamanho: 4 linhas.

Tabela 8: Dimensão Categoria

Campo	Descrição dos dados	Origem dos dados	Valores
Category Key	Chave Primária	ID gerado sequencialmente	1 - 4
Pizza Category	Tipo de categoria onde a pizza se insere		"Veggie"; "Classic"; "Supreme"; "Chicken"

8.4 Dimensão Pedido – DimOrder

A dimensão *Order* permite analisar cada pedido realizado, englobando todas as transações contidas no mesmo, fornecendo uma visão detalhada do valor financeiro e da composição dos pedidos. Esta dimensão é fulcral para entender o comportamento de compra dos clientes e para a gestão financeira das vendas.

Esta dimensão apenas tem o atributo "*Total Price*", que engloba a soma de todas as transações associadas a um pedido.

Tamanho: 21 350 linhas.

Tabela 9: Dimensão Pedido

Campo	Descrição dos dados	Origem dos dados	Valores
Order Key	Chave Primária	ID gerado sequencialmente	1 - x
Total Price	Valor total de uma transação (todas as pizzas)		9.75 - 83

9. Diagrama em estrela

Este diagrama ajuda a visualizar como se relacionam as dimensões, entre si destacando a *DimPizza e DimCategory,* e com a tabela de factos.

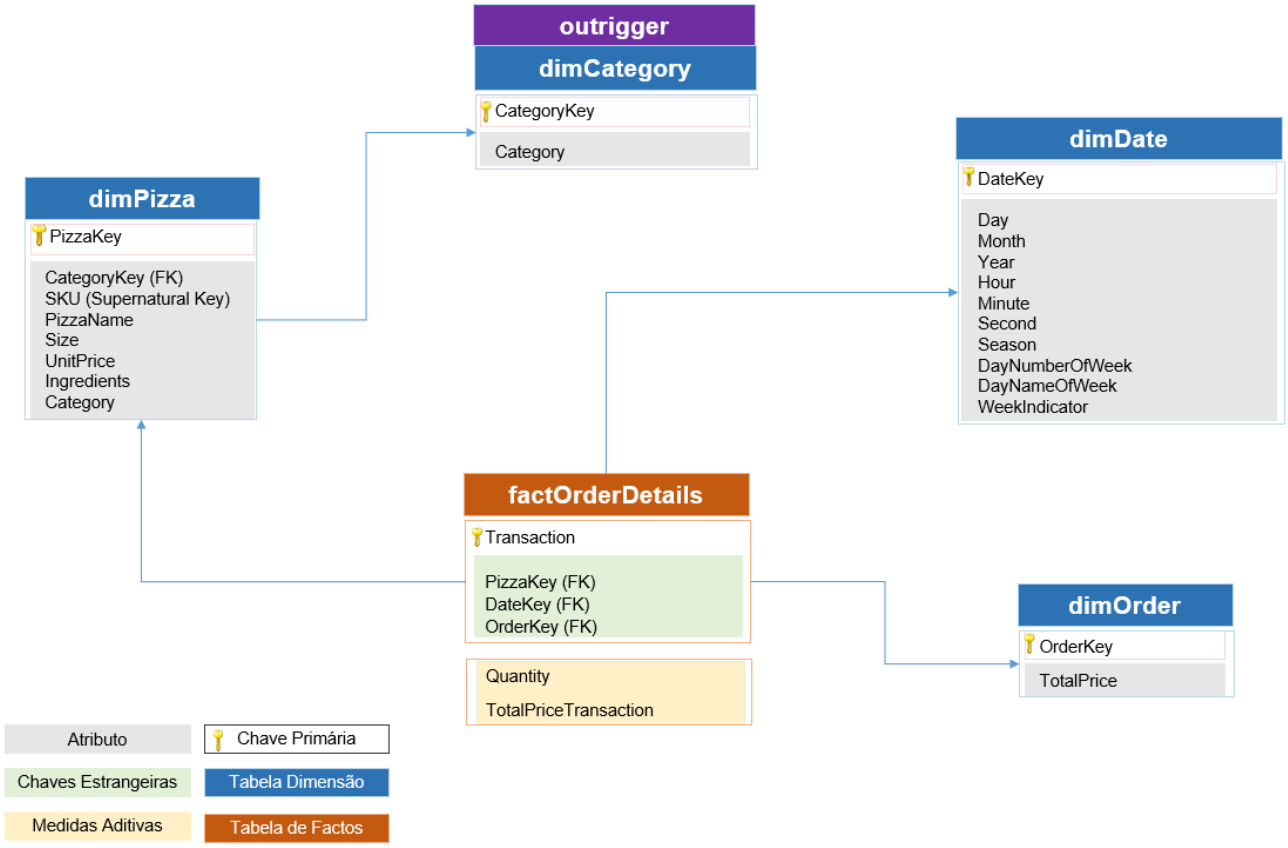

Figura 9: Diagrama em estrela

10. ETL

Num contexto computacional, um sistema ETL é constituído como um processo de três fases: *Extract, Transform, Load*. Nesta fase, pretende-se seguir este mesmo processo, extraindo os dados da fonte original, aplicando as transformações necessárias e respetiva limpeza aos mesmos, o que permitirá o carregamento da informação para uma base de dados, possibilitando a aplicação das análises

pretendidas, no nosso cenário, recorrendo à ferramenta de visualização de dados PowerBI Desktop e a bibliotecas de Python.

O ETL serve como "porta de entrada" para o sistema OLAP, sendo neste transformada e uniformizada toda a informação proveniente dos sistemas OLTP. É nesta fase que a informação é processada, sendo extraída a informação realmente relevante que irá ser armazenada no sistema OLAP, bem como efetuada a uniformização desta informação para os moldes do sistema de destino.

Todo o script foi implementado em *Python,* existindo, no entanto, statements de *SQL*. A nossa arquitetura segue um paradigma de programação orientada a objetos, a qual facilita na utilização da interface, bem como modulariza todo o processo, o que flexibiliza as possíveis futuras alterações, bem como providencia possibilidade de reutilização do código.

Cada uma das dimensões tem a sua própria classe, a qual herda atributos e métodos de uma classe abstrata, representando globalmente qualquer dimensão. Nesta classe abstrata definem-se métodos tais como a inicialização das tabelas de lookup (explicadas mais adiante), fim do processamento e processamento de uma linha do sistema operacional. Em cada uma das classes específicas, de cada dimensões, existem implementações para o processamento de linha do sistema operacional, pois este procedimento é específico a cada uma destas dimensões.

De modo a correr fluidamente todo o processo, existe apenas uma função a qual é preciso chamar para executar todo o sistema. Nesta, são instanciados os objetos relativos a cada dimensão, bem como o objeto da tabela de factos. De seguida é processada individualmente cada linha proveniente dos sistemas OLTP.

Embora conceitualmente o sistema ETL seja segmentado em três etapas distintas, achámos mais simples e intuitivo a programação destas etapas "fundidas" entre elas, isto é, a transformação e o carregamento acontecem em convivência, sendo carregada para a base de dados cada linha que é processada, em vez de processadas todas as linhas sequencialmente e posteriormente carregadas em simultâneo para a base de dados.

É também importante referir que, de modo a separar chaves únicas entre os sistemas OLTP e OLAP e evitar problemas, nenhum dos IDs provenientes do dataset original foi utilizado como Surrogate Key no sistema OLAP.

10.1 Extração

Como mencionado anteriormente, a primeira fase deste sistema passa pela extração de dados. No presente projeto, este processo passou pelo download de um dataset

proveniente da plataforma Kaggle que, após sofrer uma análise superficial, mostrou-se adequado ao propósito do projeto. Refere-se ainda que o formato original dos dados utilizados encontra-se em Excel.

10.2 Transformação

Esta é a fase mais crítica e importante do ETL, onde são transformados todos os dados provenientes dos sistemas OLTP.

Realizou-se a limpeza destes dados, ou seja, resolução de erros, duplicados, exceções, valores nulos, outliers, bem como a preparação/uniformização para futura introdução no sistema OLAP.

Este processo foi codificado e está disponível no jupyter notebook, entregue juntamente a este relatório, na célula identificada com o título "Dimensões".

É também importante mencionar que, no conjunto de dados original, sendo apenas uma fonte, em formato Excel, não existia uma distinção de tabelas, estando todos os campos incorporados neste ficheiro. Estas ficaram definidas pelo grupo nas fases anteriores do presente relatório, sendo processadas no código mencionado.

10.2.1 Tratamento de dados para a dimensão Categoria

Responsabilidade: Tratamento e uniformização dos dados relativos à categoria da pizza. No âmbito deste projeto, uma pizza pode ser classificada por quatro categorias diferentes: Classic, Veggie, Supreme e Chicken. Tendo em conta que esta coluna, no dataset original, está "limpa", isto é, não apresenta erros, não foi necessário nenhum procedimento extra para a sua utilização.

Input

- Campo *pizza_category*, do dataset original

Output

- Tabela uniformizada com cada uma das categorias existentes

10.2.2 Tratamento de dados para a dimensão Pizza

Responsabilidade: Tratamento e uniformização dos dados relativos a cada pizza, incluindo *SKU (Stock Keeping Unit), preço unitário, tamanho e ingredientes.* Os *ingredientes*, originalmente numa string separados por vírgulas, foram transformados para um array de PostgreSQL, o que permite maior flexibilidade nas futuras análises relativas a estes dados.

Input

- Os campos relativos a atributos da pizza:
    - preço unitário
    - tamanho
    - ingredientes
    - nome

Output

- Tabela uniformizada com cada pizza existente, com os atributos descritos no Capítulo 9, bem como atributos do tipo 2 de SGD (Slowly Changing Dimensions).

10.2.3 Tratamento de dados para a dimensão Data

Responsabilidade: Tratamento e uniformização dos dados relativos aos campos da data e hora. Nesta dimensão, a separação dos atributos da data e da hora em várias colunas facilita futuras análises, pelo que foram extraídos atributos específicos do ficheiro fonte, tais como dia do mês, mês, ano, hora, minuto, segundo, número do dia da semana, nome do dia da semana e indicador de fim de semana.

Input

- Os dois campos associados a tempo:
    - *order_date* no formato dd/mm/aaaa
    - *order_time* no formato hh:mm:ss

Output

- Dimensão com campos uniformizados e mais granulares:
    - Day
    - Month
    - Year
    - Hour
    - Minute
    - Second
    - Season
    - DayNumberOfWeek
    - DayNameOfWeek
    - WeekendIndicator

10.2.4 Tratamento de dados para a dimensão Order

Responsabilidade: Tratamento e uniformização dos dados relativos ao pedido. Tendo em conta que estes dados não são fornecidos explicitamente no dataset original, para esta dimensão é calculado o valor total de cada pedido, através da agregação das transações relativas a cada um destes pedidos.

Input

- *Total_Price,* campo representativo do preço total de uma transação, bem como preços gastos nas transações ligadas ao mesmo pedido.

Output

- Dimensão com campos uniformizados:
    - Total (este é relativo à agregação e soma dos valores das transações relativos a cada pedido, isto é, preço total do pedido)

10.2.5 Tratamento de dados para a tabela de factos

Responsabilidade: Tratamento e uniformização dos dados relativos a cada transação de cada pedido.

Input

- Surrogate Keys de cada tabela de dimensão correspondente aos dados da transação, bem como o valor total da mesma

Output

- Tabela de factos com ligações a cada dimensão e o valor total da transação

10.3 Carregamento

O carregamento dos dados, anteriormente extraídos e transformados, corresponde à terceira etapa do processo ETL. Nesta fase, os dados preparados são carregados para uma base de dados para análise. O procedimento pode ser consultado na última célula do notebook "ipai-ETL" disponibilizado, onde se estabeleceu a conexão com a base de dados PostgreSQL. De modo a ter um ambiente dinâmico no qual os quatro elementos do grupo pudessem alterar valores na base de dados e manter uma versão coerente entre todos, foi efetuada uma ligação para o servidor disponibilizado pelos docentes nas aulas práticas (appserver-01.fc.di.ul.pt). Deste modo, todos os elementos do grupo acedem à mesma versão da base de dados, mitigando erros e inconsistências.

11. Flow-Chart

Na seguinte figura é possível visualizar o *flow* de informação dentro do sistema ETL, com todas as suas fases inerentes.

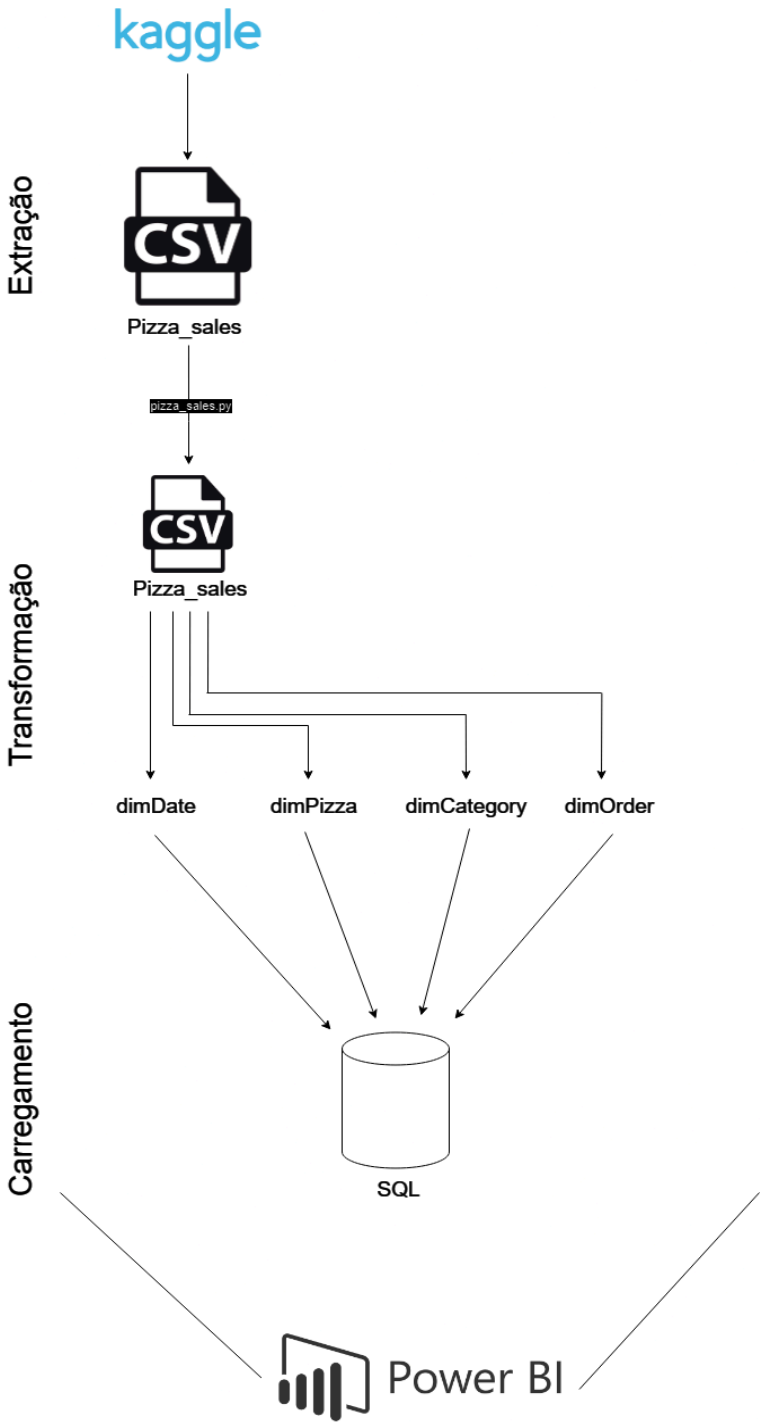

Figura 10: Flow-Chart.

12. Data Cube

Recorreu-se ao software PowerBI para a criação e respetiva visualização do cubo de dados. Foi possível, através do PostgreSQL, automatizar o processo de carregamento das dimensões, tabelas de factos e as suas ligações diretamente para o PowerBI.

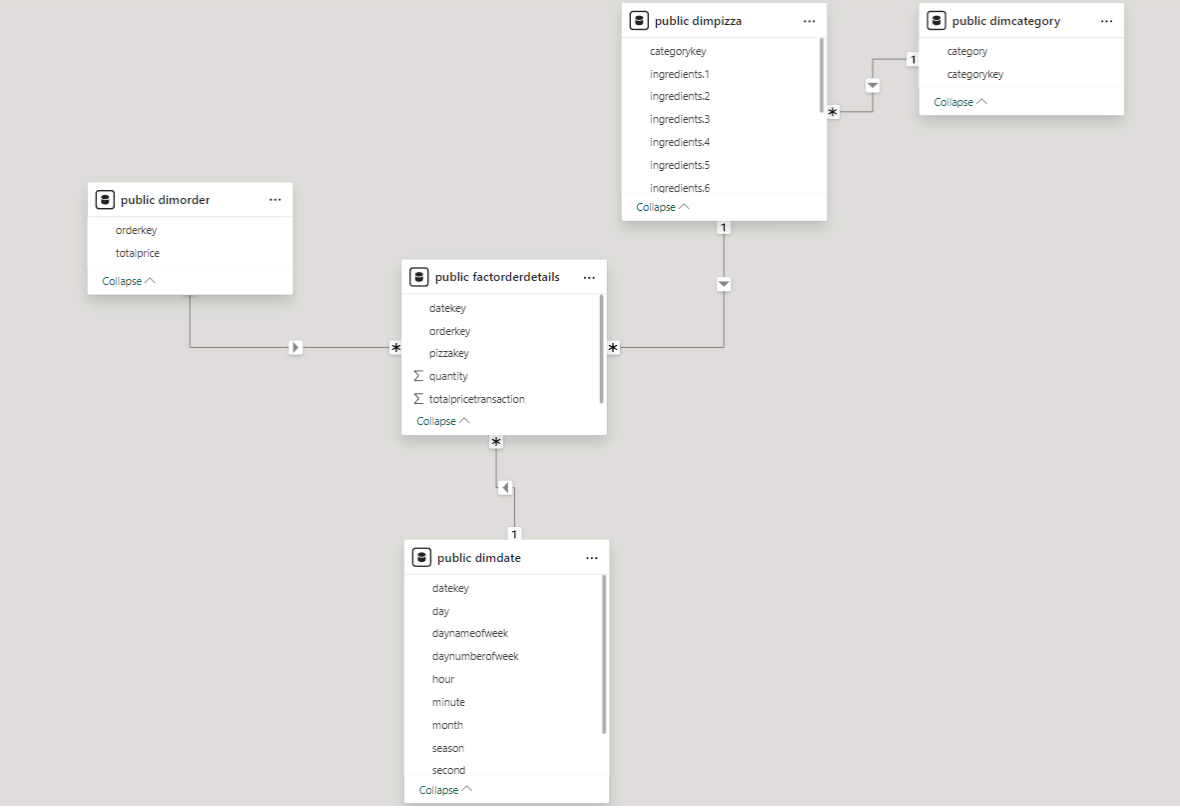

Figura 11: Data Cube obtido através do PowerBI.

13. Respostas Analíticas

13.1 Questão analitica 1

"Como é que a composição de ingredientes de cada tipo de pizza influencia a sua popularidade e o valor médio de vendas? "

Numa análise inicial, começámos por tentar entender quais são as pizzas mais populares, bem como o preço médio das mesmas. Para tal, recorreram-se a gráficos de funil. De seguida, através de um pie chart, foi analisada a soma do preço das transações por pizza. Algo também interessante de analisar foi a popularidade de alguns ingredientes em relação às pizzas mais vendidas. Para finalizar, foi utilizado um line chart, no qual é comparada a soma total de pizzas vendidas com o preço total das transações por pizza.

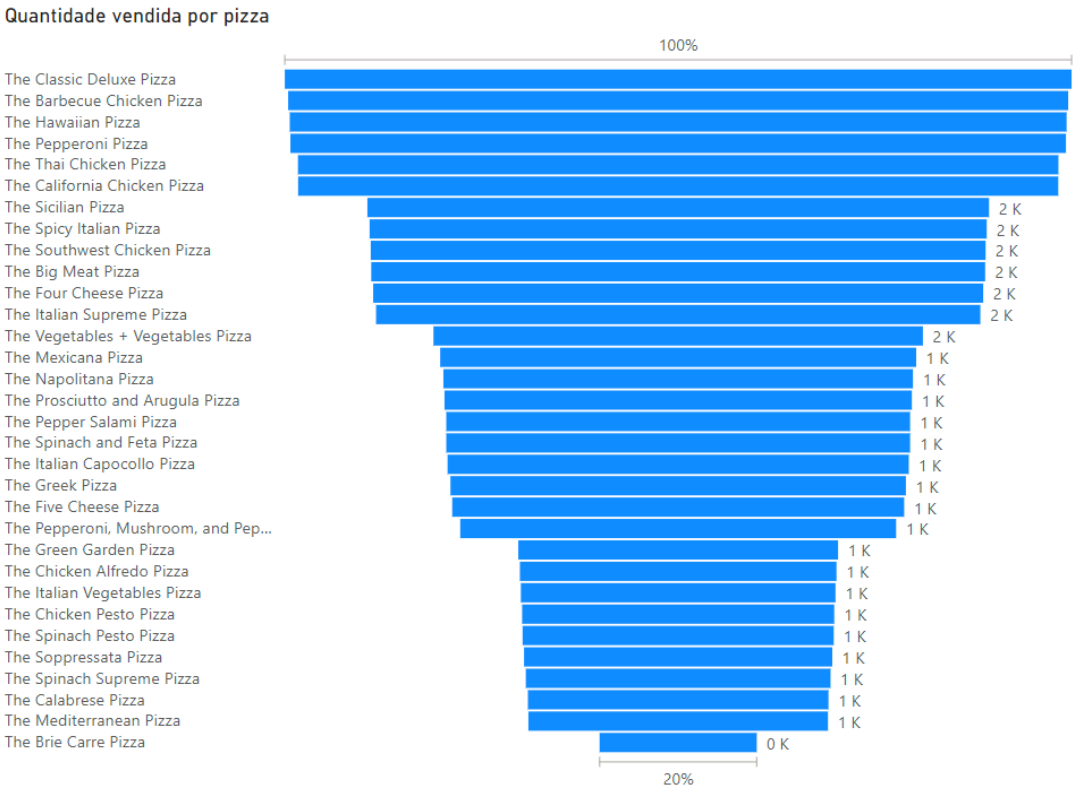

Figura 12: Quantidade vendida por pizza

A Figura anterior permite analisar quais são as pizzas mais solicitadas, bem como as menos.

No topo do gráfico, com o maior número de vendas, apresentam-se pizzas como “The Classic Deluxe Pizza”, “The Barbecue Chicken Pizza” e “The Hawaiian Pizza”, enquanto com o menor número de vendas tem-se “The MediterraneanPizza“ e “The Brie Carre Pizza”.

Este gráfico permite ter uma visão de alto nível acerca do que o restaurante mais vende, o que permitiria, por exemplo, retirar pizzas menos vendidas do cardápio, ou mesmo mudar a disposição dos nomes no menu de modo a apresentar primeiro as pizzas menos vendidas.

A análise desta informação em conjunto com a análise dos ingredientes mais vendidos poderá fornecer perspetivas ainda mais profundas e interessantes.

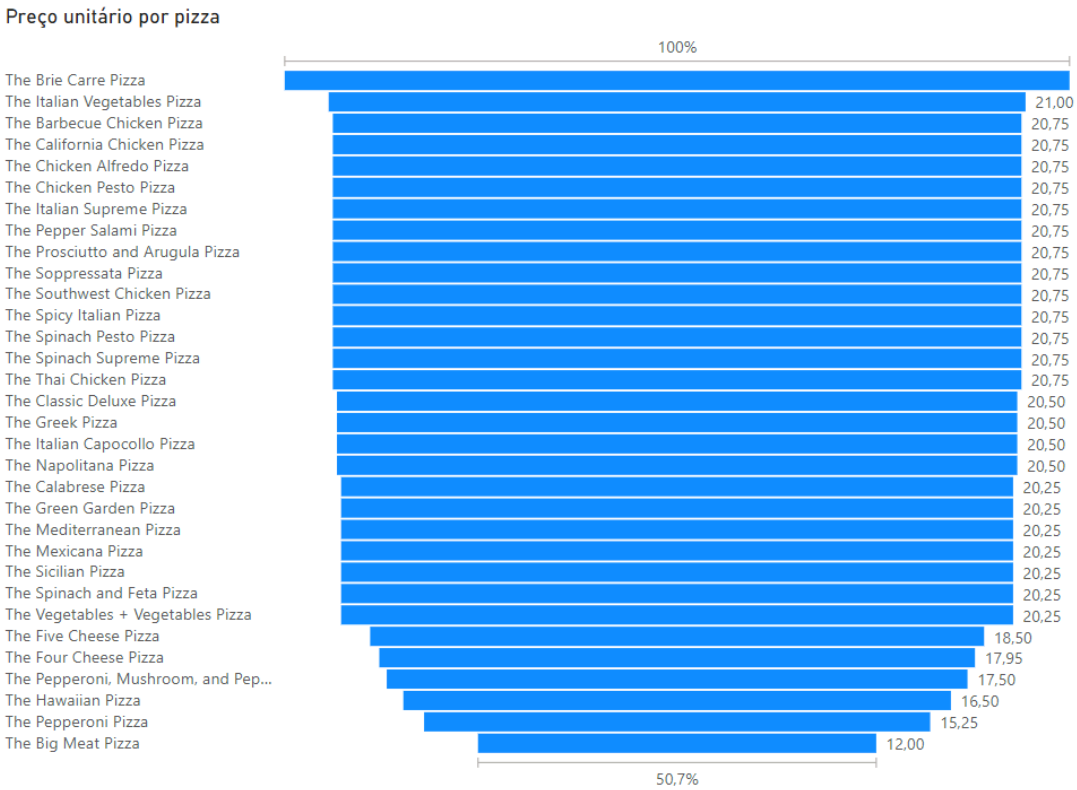

Figura 13: Preço Unitário por pizza

Seguindo para a análise do preço unitário de cada pizza, feita através do gráfico de funil, conseguimos perceber que a pizza com o valor mais caro é a “The Brie Carre Pizza” com o custo de 23,65€, seguida da “The Italian Vegetables Pizza” com um custo de 21€. Em contrapartida, a pizza mais barata é a “The Big Meat Pizza” com um custo de 12€. No gráfico da figura X (figura anterior), foi verificado que a pizza com maior quantidade de pedidos é a “The Classic Deluxe Pizza” que se encontra numa das pizzas mais caras da pizzaria. Isto poderá indicar que, apesar de existirem ofertas de pizzas mais baratas, há uma grande preferência para a “The Classic Deluxe Pizza”, pelo que entender o porquê do sucesso desta pizza seja importante (no entanto, esta análise não nos é possível concretizar sem dados mais específicos, tais como feedback dos clientes).

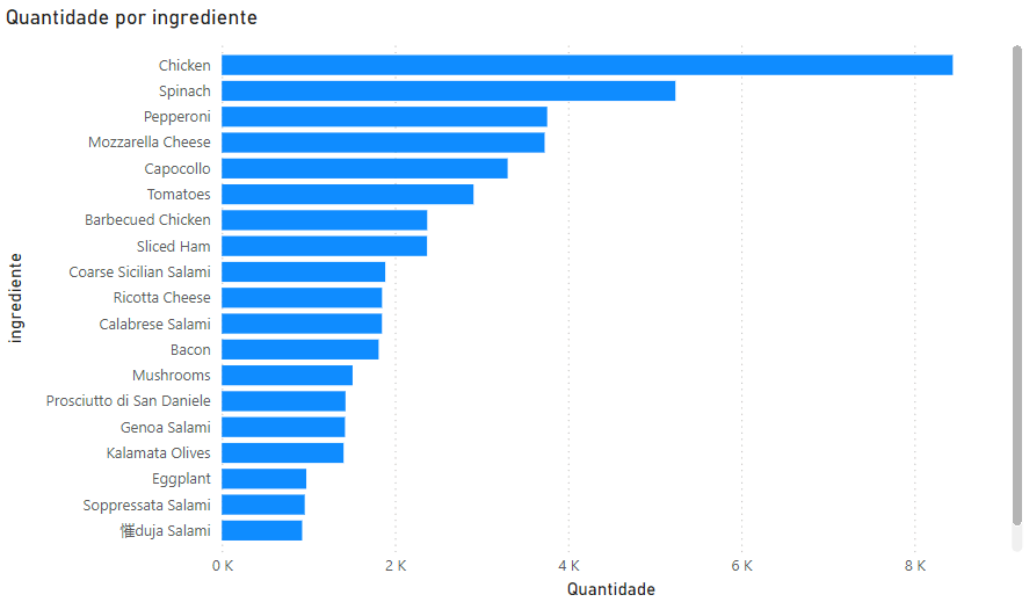

Figura 14: Quantidade por ingrediente.

Para analisar de forma mais detalhada a quantidade vendida relativamente ao preço de cada pizza, foi utilizado novamente um funil. Começando por analisar os ingredientes, salta à primeira vista que "Chicken" é o ingrediente mais utilizado nas pizzas consumidas. Estes resultados são expectáveis, tendo em conta que as pizzas mais vendidas têm este ingrediente

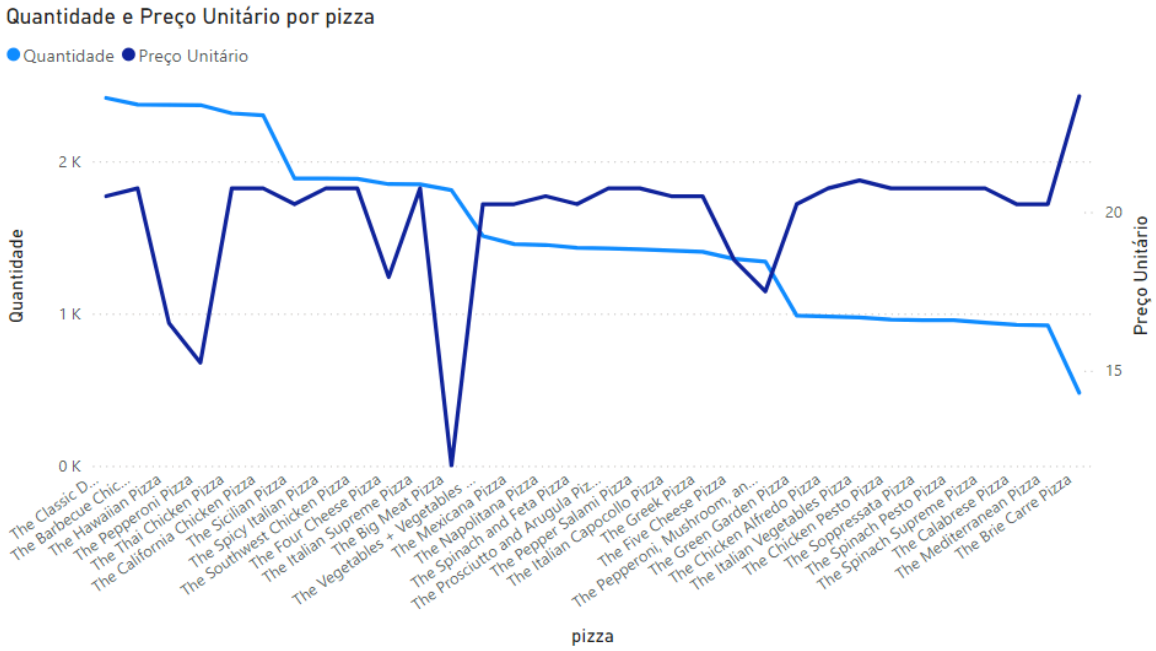

Figura 15: Quantidade e Preço Unitário por pizza.

Por fim, através da análise da figura anterior com o line chart, composto pela soma das quantidades e pela pelo valor de cada pizza, conseguimos também concluir que as

pizzas mais vendidas nem sempre são as mais baratas, como, por exemplo, “The Classic Deluxe Pizza”. No entanto, a pizza menos vendidas é, efetivamente, a pizza mais cara disponível no restaurante.

13.2 Questão analítica 2

“Quais são os dias e horas mais movimentados do restaurante?”

Dias mais movimentados

Começou-se por analisar quais são os dias mais movimentados no restaurante ao longo da semana. Para tal, foram utilizados um line chart, um donut chart e um treemap.

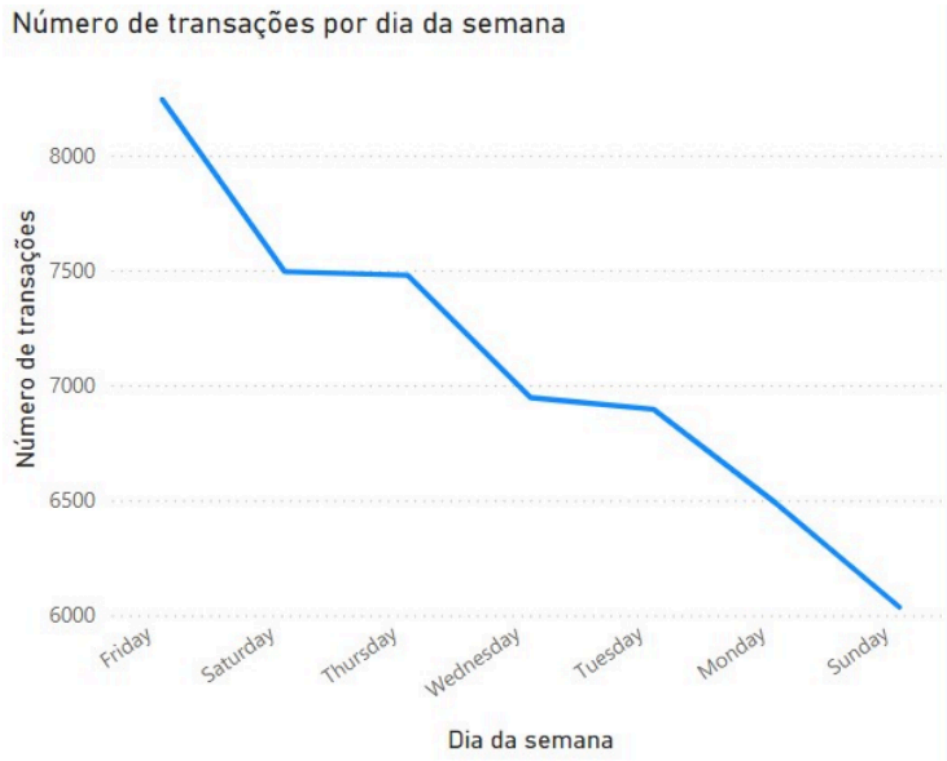

Figura 16: Número de transações por dia da semana.

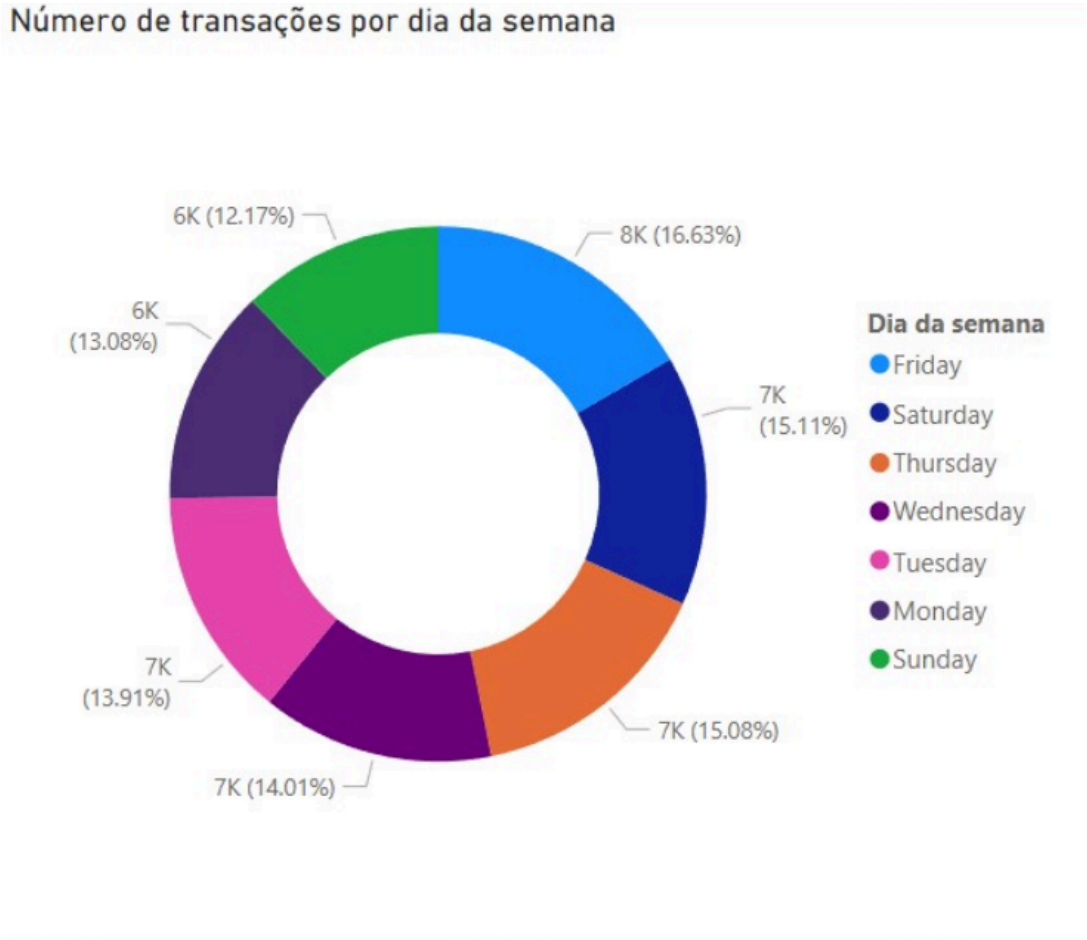

Figura 17: Número de transações por dia da semana.

Figura 18: Número transações por dia da semana.

Através da observação e análise cuidada das três figuras apresentadas, é possível entender que os dias mais movimentados no restaurante são a sexta-feira e o sábado. Estes resultados são os expectáveis, tendo em conta que estes dias são o começo e o primeiro dia do fim de semana, nos quais é comum existir maior disponibilidade temporal da população para atividades sociais como saídas noturnas, nas quais estão tipicamente incluídas as refeições em estabelecimentos de restauração.

Os dias menos movimentados são o domingo e a segunda-feira. Estes também são resultados esperados, tendo em conta que são o "começo" da semana, e a maioria das pessoas já gastou mais dinheiro no início/durante o fim de semana. Culturalmente, também é muito comum existirem almoços ao domingo em que as famílias se reúnem, pelo que este pode ser outro fator impactante para explicar estes valores

Horas mais movimentadas

De modo a perceber quais são as horas mais movimentadas no restaurante, utilizaram-se um ribbon chart, um donut chart e um treemap, pelo facto de nos parecerem as visualizações mais apropriadas para o contexto.

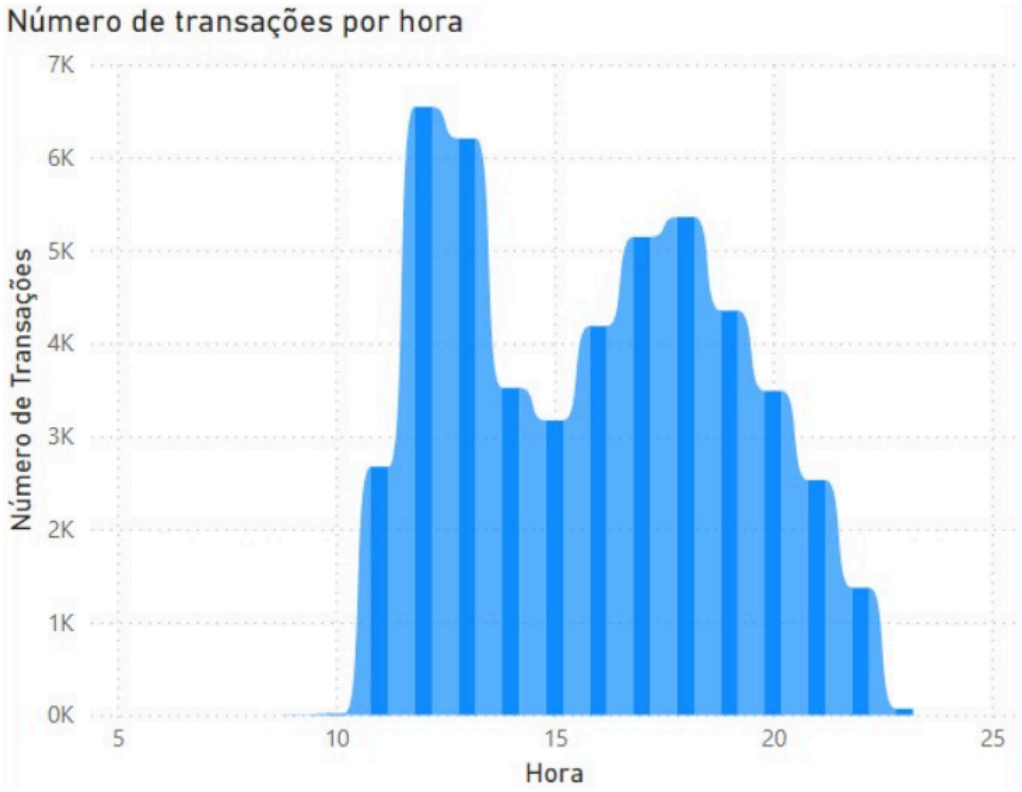

Figura 19: Número de transações por hora.

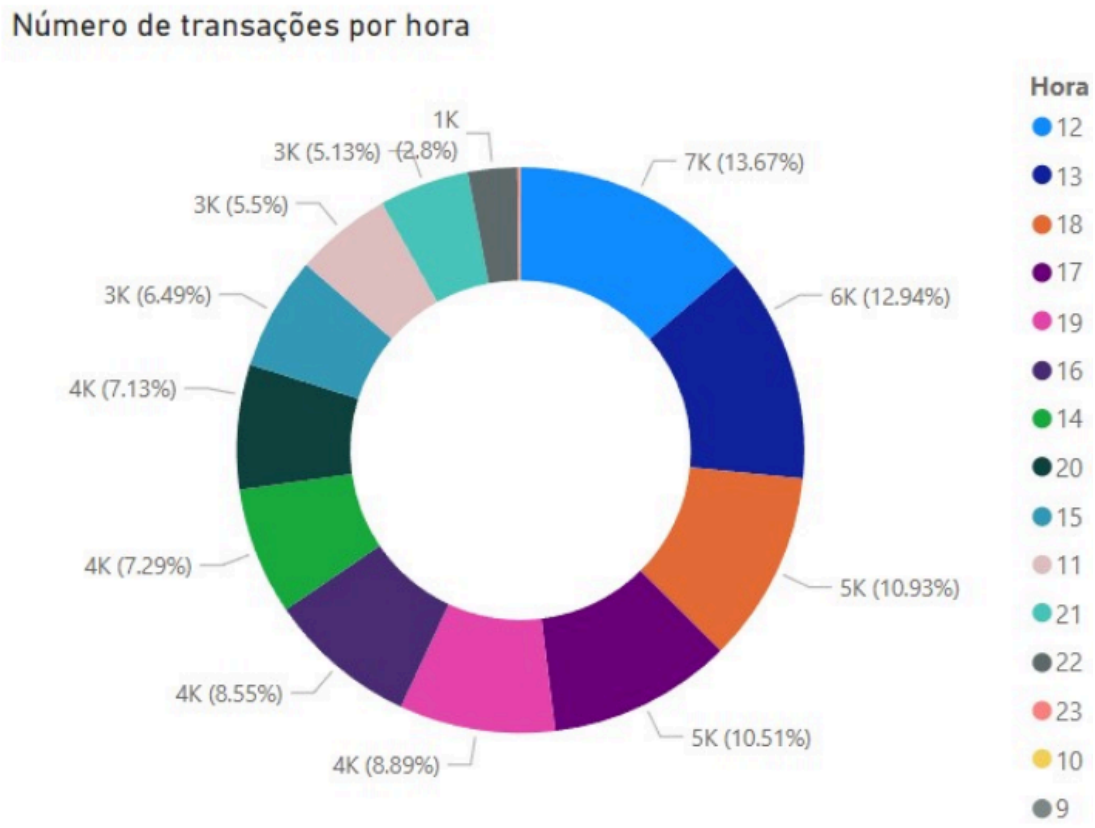

Figura 20: Número de transações por hora.

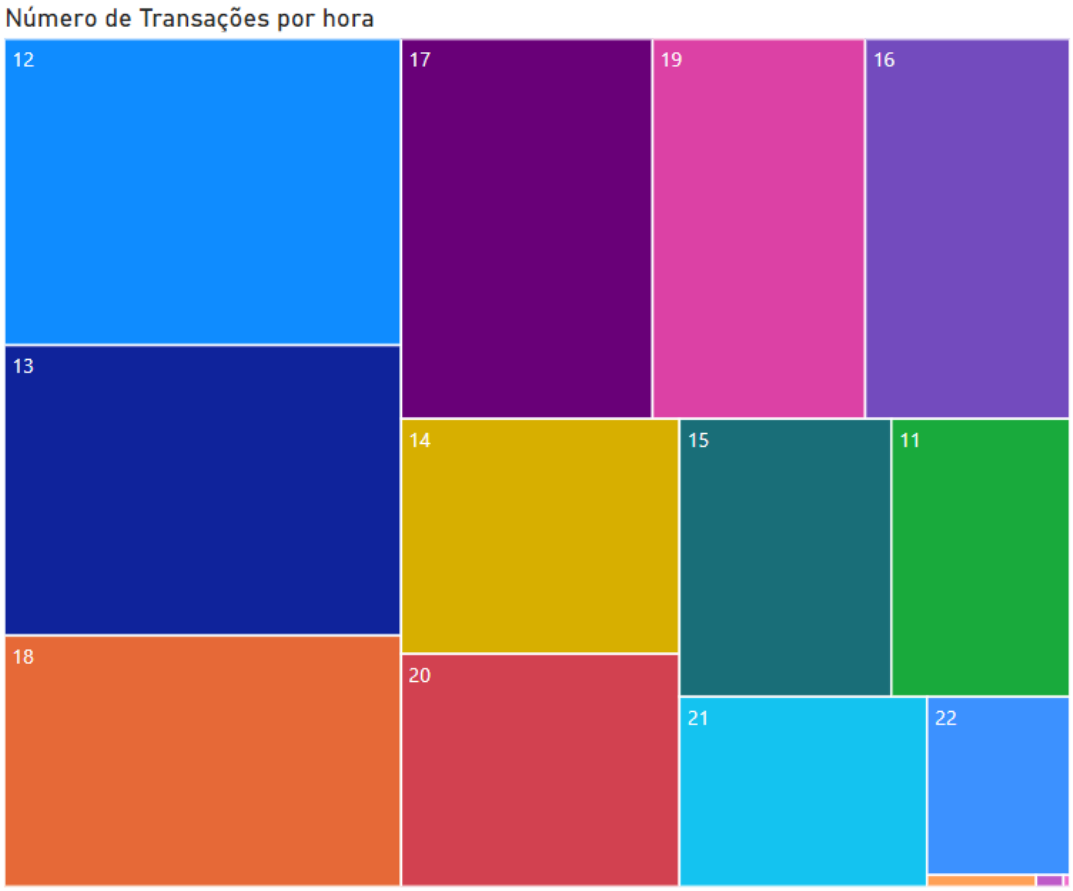

Figura 21: Número de transações por hora.

Figura 22: Número de transações por hora aos fins de semana.

Através da análise das figuras apresentadas é possível perceber que as horas mais movimentadas concentram-se entre as 12h-13h e as 17h-19h. Tendo em conta que, tipicamente, para a maioria da população, a hora de almoço varia entre as 12h-14h, tanto culturalmente como em termos de horários empresariais, e a hora do jantar entre as 19h-21h, estes resultados são interessantes.

Relativamente ao almoço, os valores obtidos são "normais", dado que entre as 12h-13h são as horas habituais de almoço da grande maioria da população. O mais interessante passa pelas horas mais movimentadas na chamada "hora de jantar". O expectável seria vermos valores entre as 19h-21h, como referido anteriormente. No entanto, os dados dizem-nos que é entre as 17h-19h. Estes valores podem ser possivelmente explicados por motivos culturais, dado que as horas de refeições variam consoante a cultura e a localização geográfica, pelo que os valores que seriam expectáveis para nós (sendo "nós", a população portuguesa, no geral), podem não ser iguais aos do país onde o restaurante que estamos a analisar se encontra.

Durante o fim de semana as horas mais movimentadas alteram-se um pouco. É possível observar que a tendência deixam de ser os almoços, e passam a ser os jantares. Dentro das "horas de almoço", é possível entender que os clientes vão mais tarde, com um número de vendas às 12h a ser bastante inferior no fim de semana.

Análises complementares

De modo a apresentar uma análise mais profunda e robusta, fomos também analisar os meses mais movimentados, as estações do ano mais movimentadas, bem como entender a diferença de movimento do restaurante entre os dias de semana e o fim de semana.

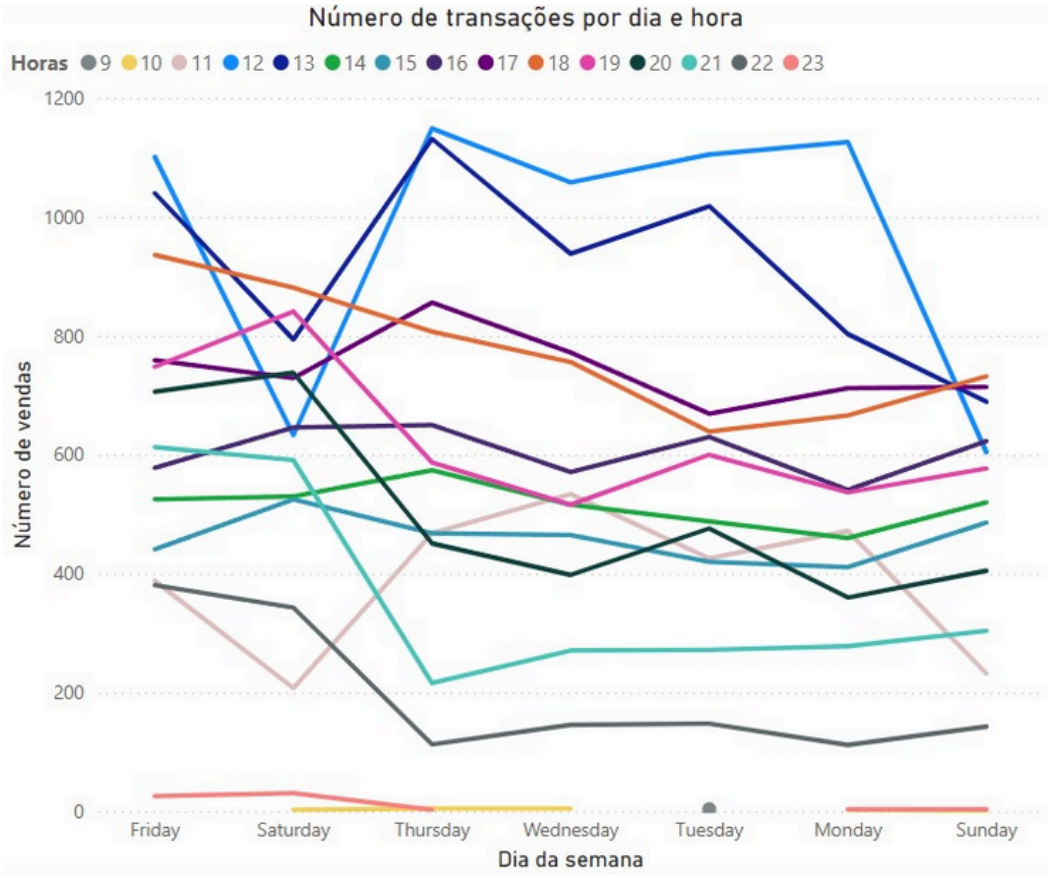

Figura 23: Número de transações por dia e hora.

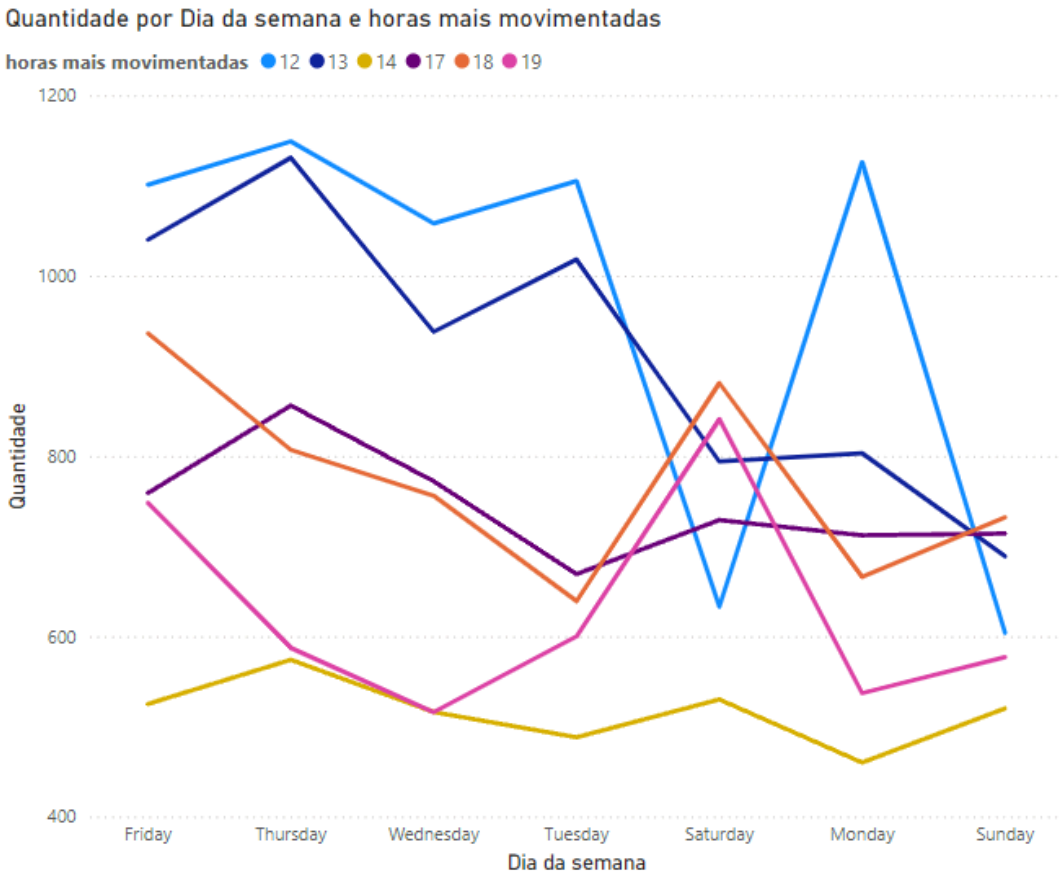

Figura 14: Quantidade por dia da semana e horas mais movimentadas.

É possível observar que os dados apresentados na figura anterior coincidem com a análise feita anteriormente. No entanto, aqui pode ser verificado que, por exemplo, ao sábado, embora seja um dos dias mais movimentados, esse movimento está concentrado nas horas de jantar, não existindo para as horas de almoço tantas vendas como em dias da semana como, nomeadamente, a quarta-feira.

Figura 25: Número de transações por mês

Através da análise do anterior gráfico de funil percebe-se que o mês mais movimentado do ano é julho. Este resultado pode ser explicado pelo facto de que este é um mês do verão onde uma grande parte da população tem férias, tendo então maior disponibilidade temporal para fazer refeições em restaurantes, bem como maior capacidade financeira.

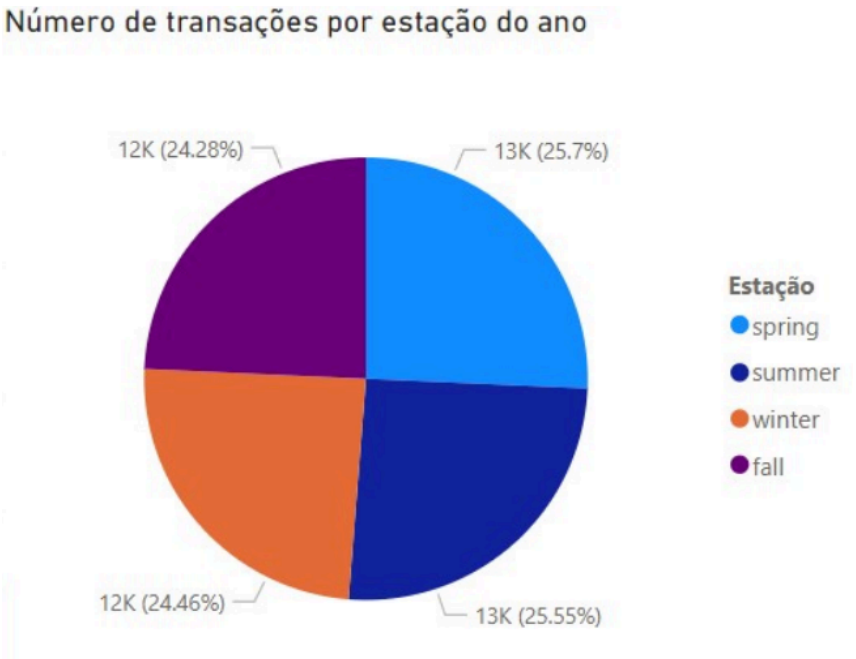

Figura 26: Número de transações por estação do ano.

Na figura anterior observa-se a quantidade de transações por cada estação do ano. É possível entender que, embora o verão apresente um número de transações ligeiramente superior às outras estações, este é módico, pelo que, à partida, não é possível estabelecer qualquer ligação entre a quantidade de vendas e as estações do ano.

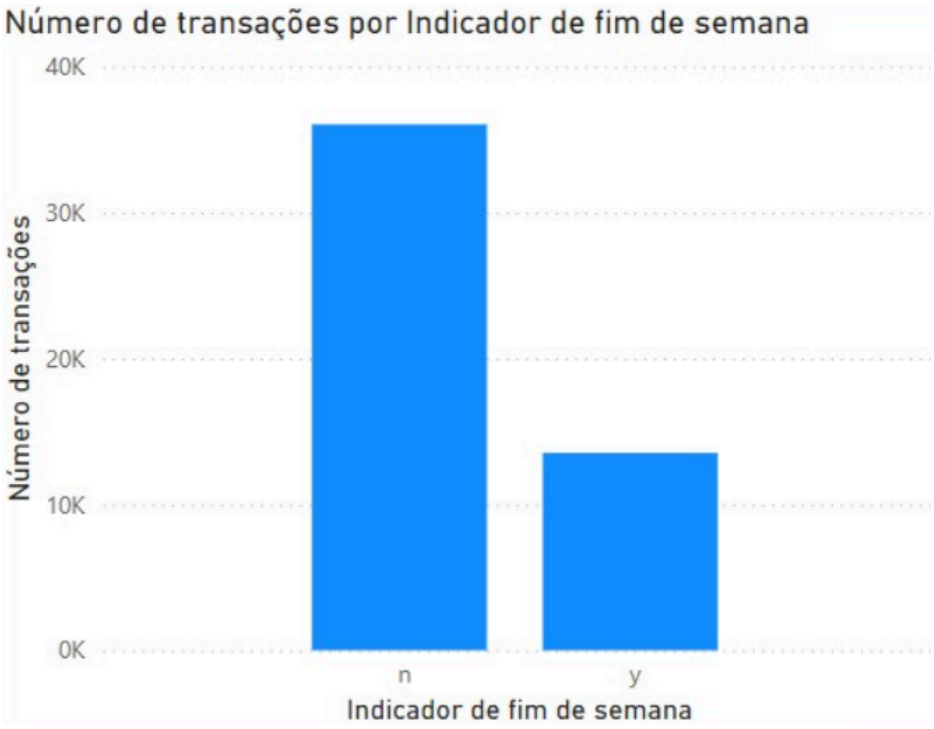

Figura 27: Número de transações por indicador de fim de semana,

Por fim, utilizou-se um histograma (gráfico de barras) com vista a analisar a diferença entre o movimento dos dias úteis e do fim de semana.

Número de Transações (aproximado)	Número de dias	Rácio
35000	5	7000
13500	2	6750

Através do cálculo do rácio *número de transações/dia*, apresentado na tabela anterior, entende-se que, por dia, existe tipicamente mais movimento durante a semana do que no fim de semana.

13.3 Questão analítica 3

"De que maneira está a pizzaria a utilizar a sua capacidade de lotação?"

É extremamente útil a qualquer espaço de restauração entender de que maneira o seu espaço é utilizado, de modo a otimizar a utilização dos recursos disponíveis, assim como melhorar a experiência do cliente. Estes aspetos podem, eventualmente, ser a diferença entre um negócio de sucesso e o fecho de portas.

Como já mencionado ao longo deste relatório, a pizzaria Plato´s conta com um total de 15 mesas e 60 cadeiras atualmente. Pretende-se, com esta análise, entender a melhor forma de organizar estes materiais no espaço disponível e perceber se são ou não suficientes.

De modo a tornar esta análise possível (e também repetindo o que já fora mencionado na Capítulo 7), considera-se que uma pessoa come uma pizza, portanto cada pizza pedida corresponde a um indivíduo. De modo a perceber quantos "grupos" existem no restaurante, utiliza-se o campo *order_id* (identifica um pedido completo), que identifica um conjunto de transações, isto é, um conjunto de pessoas

Recorrendo ao ficheiro de dados original para um breve exemplo, é possível verificar, na figura seguinte, que existem duas mesas ocupadas no restaurante, divididas num grupo 5 de indivíduos e outro indivíduo sozinho (é possível entender a quantidade de indivíduos na mesa através da quantidade pedida em cada transação).

order_details_id	order_id	pizza_id	quantity
1	1	hawaiian_m	1
2	2	classic_dlx_m	1
3	2	five_cheese_l	1
4	2	ital_supr_l	1
5	2	mexicana_m	1
6	2	thai_ckn_l	1

Acrescenta-se ainda que, devido à ausência de informação, considera-se que todas as mesas presentes no restaurante têm o mesmo tamanho.

13.3.1 Afluência de pessoas numa semana específica

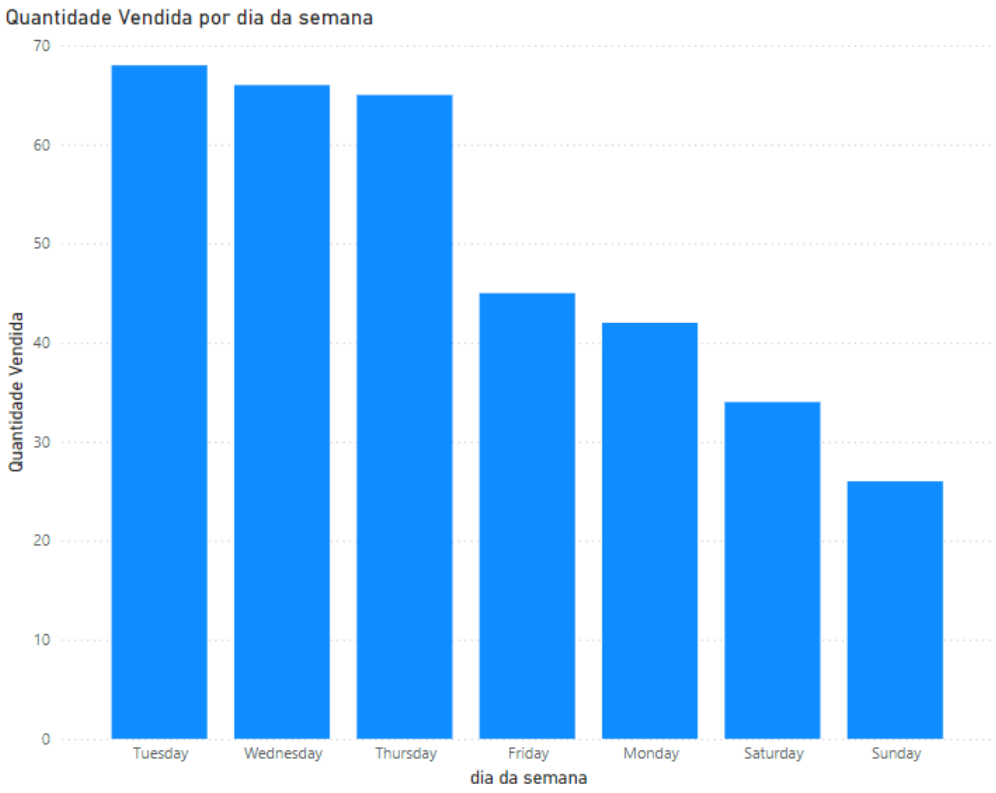

Figura 27: Quantidade vendida por dia da semana.

Através da figura anterior é possível observar a quantidade de pizzas vendidas, por dia da semana, entre os dias 1 e 7 de janeiro de 2015, entre as 12 e as 14 horas. É possível entender que, por exemplo, houveram cerca de 67 vendas na terça-feira, o que representa um número superior ao número de lugares disponíveis, demonstrando um bom uso do espaço, através da rotação de clientes. No entanto, existe espaço para melhoramento, como é possível observar por dias como domingo ou sábado. Estes dados reforçam as análises já apresentadas anteriormente, relativamente ao movimento por dia da semana.

14. Conclusão

Na primeira etapa do projeto colocaram-se esforços na identificação de fontes de dados pertinentes para a realização do projeto. Foi estabelecido um processo de negócio, a partir

do qual foram tratados e organizados os dados recolhidos. Procedeu-se posteriormente à alteração e adição de campos, assim como a definição de tabelas, as quais podem ser consultadas através do diagrama apresentado no Capítulo 4. Por fim, em relação ao processo de negócio estabelecido, foram colocadas três questões analíticas para investigação.

Na segunda etapa do projeto começámos por estruturar uma ideia das dimensões adequadas à utilização deste projeto, partindo numa fase inicial de três dimensões. De seguida, procedemos à consequente criação da tabela de factos e definição do grão, onde se integraram as dimensões idealizadas, no entanto, apercebemo-nos das vantagens analíticas perante o modelo de negócio proposto, de se adicionar uma nova dimensão em forma de outrigger, a partir da dimensão Pizza, sendo esta nova dimensão denominada por Categoria. Na tabela de factos descrevem-se também as medidas aditivas *Quantity* e *TotalPriceTransaction*. Por fim, disponibiliza-se o diagrama em estrela, o qual permite de forma clara a identificação das ligações entre tabela de factos e dimensões, assim como características dos seus campos.

Na terceira e última fase do projeto foi implementado o sistema de ETL com base no planeamento realizado nas duas anteriores etapas. Para tal foi processado o código num notebook, que inclui as fases de extração, transformação e carregamento, o mesmo disponibiliza-se juntamente com esta entrega. Todo o processo é realizado sem qualquer constrangimento, carregando a base de dados para o servidor "appsever-01.di.fc.ul.pt", o que facilitou a implementação por parte da equipa.

Procurou-se responder de maneira adequada às perguntas analíticas definidas anteriormente, recorrendo à ferramenta de software PowerBI. Foram encontradas algumas dificuldades na questão 3, a qual não nos foi possível realizar uma análise mais aprofundada por questões de tempo.

Menciona-se por fim que todo o nosso processo de código se encontra no notebook mencionado, devidamente comentado e organizado, facilitando assim o seu entendimento. Optámos por o realizar desta forma face ao contexto do nosso sistema ao nivel do tamanho de dados e o seu processamento.