Relatório - Tecnologias Escaláveis para a Análise de Dados

João Bragança (8190555)
Pedro Afonso (8090457)
José Fernandes (8190239)

Grupo 3

Escola Superior de Tecnologia e Gestão P.Porto

May 7, 2023

Contents

1 Introdução

1.1 Enquadramento

No âmbito da Unidade Curricular *Tecnologias Escaláveis para Análise de Dados* (TEAD), inserida no primeiro ano do mestrado em Engenharia Informática da Escola Superior de Tecnologia e Gestão, lecionada pelo professor Davide Carneiro, foi-nos proposto simular de forma realista o ciclo de vida de um projeto de análise de dados. Para isso será realizado um projeto em conjunto com uma empresa nacional do ramo da hotelaria, a e-GDS cujo âmbito é a previsão de reservas. Neste âmbito é pretendido o seguinte:

- Fazer a análise do dados fornecidos pela e-GDS;
- Identificar fontes de dados externas que possam ser relevantes para o problema;
- Das diferentes fontes de dados analisar de forma critica a sua relevância para o problema;
- Criar e validar novas variáveis que possam enriquecer o *dataset* fornecido;
- Treinar e avaliar modelos de Machine Learning com vista à resolução do problema em concreto.

1.2 Caso de Estudo

A e-GDS é uma empresa que oferece soluções tecnológicas para a indústria hoteleira, permitindo a automatização de operações de gestão de hotéis. Com o passar dos anos foram recolhendo vários dados de reservas feitas por hóspedes dos seus clientes. Com este trabalho pretende-se desenvolver um modelo de previsão de ocupação/afluência ao longo do tempo para os diferentes hotéis cujos dados foram disponibilizados, juntamente com dados de fontes externas que sejam consideradas relevantes para a resolução do problema.

1.3 Fases do trabalho

Identificação de fontes de dados externas

O trabalho será desenvolvido em várias fases, sendo a primeira a identificação de fontes de dados externas e públicas. O objetivo passa por recolher dados que possam ser úteis para a previsão de reservas dos hotéis. Essas fontes de dados podem incluir, informações sobre eventos locais (como concertos, feiras e conferências), condições meteorológicas, tendências de viagem, entre outras.

Integração de dados

Aos dados fornecidos pela e-GDS serão integrados dados obtidos mediante fontes públicas. Além disso, é necessário juntar os diferentes conjuntos de dados fornecidos entre si (uma vez que estão em tabelas diferentes), assim como os dados externos. Esta atividade envolverá a correspondência de colunas comuns entre os diferentes conjuntos de dados e a criação de novas colunas derivadas das fontes de dados obtidas.

Análise exploratória de dados

Após a integração dos dados, será feita uma análise exploratória de dados, dando uma melhor perspetiva sobre os valores que as variáveis podem tomar e como estas se relacionam entre si.

Aquisição e tratamento de dados

No final da integração, é necessário tratar os dados resultantes da junção dos diferentes *datasets*, uma vez que, não estão prontos a ser integrados numa solução de machine learning. Existem dados em falta, dados sem significado e dados com o mesmo significado representados de formas diferentes. Para resolver o problema técnicas de limpeza de dados serão aplicadas, como por exemplo, a eliminação, substituição, normalização, entre outras.

No final do tratamento, é possível enriquecer o *dataset* com novos atributos, atributos estes que podem ser derivados dos já existentes aplicando técnicas de *Feature Engineering*.

Desenvolvimento de modelo de previsão

Já com o *dataset* completo serão aplicadas técnicas de *machine learning* que permitirão obter previsões de reservas em unidades hoteleiras.

Por fim, serão treinados e avaliados vários modelos de forma a validar a sua utilizada e eficácia na resolução do problema proposto.

1.4 Descrição do Problema

O problema abordado neste trabalho é a previsão de ocupação/afluência ao longo do tempo em hotéis. Embora a empresa e-GDS disponibilize um *dataset* com dados de reservas efetuadas pelos seus clientes, o objetivo do trabalho vai além da utilização desses dados. É necessário identificar fontes de dados externas e públicas que possam ser indicadores importantes para a previsão de

afluência e integrá-las com os dados fornecidos pela e-GDS para análise conjunta. Assim, o desafio é adquirir, tratar e integrar essas fontes de dados com os dados existentes para melhorar a previsão de ocupação/afluência em hotéis. Além da previsão de ocupação/afluência ao longo do tempo, outro problema secundário é a previsão do cancelamento de reservas em hotéis. É importante que o hotel possa prever com antecedência quais reservas que poderão ser canceladas, permitindo que possam atuar sobre essa informação atempadamente. Dessa forma, o desafio é identificar os indicadores relevantes para a previsão de cancelamentos de reservas, bem como utilizar técnicas de análise de dados e modelagem para fazer previsões. A resolução desse problema pode trazer benefícios para o hotel, como a possibilidade de otimizar a gestão de disponibilidade de quartos e maximizar a ocupação.

2 Revisão da Literatura

2.1 Influências nas reservas de hotéis

A influência nas reservas de hotéis é um tema com um impacto relevante para a indústria hoteleira, uma vez que, a ocupação é um fator-chave para o sucesso financeiro de um hotel. Diversos fatores podem influenciar a reserva de um hotel, tais como: a satisfação do cliente, as avaliações online, os eventos, feriados, fatores meteorológicos, períodos de férias, preços entre outros. Lee e Jeon (2012)[4] mostram que a satisfação dos clientes está diretamente relacionada à qualidade dos serviços e instalações do hotel, à interação com os funcionários, à limpeza e conforto dos quartos, e ao valor percebido relativamente ao preço. Além disso, eles destacam que a intenção de retorno é influenciada pela satisfação anterior e pelo compromisso com a marca do hotel. Para Xu e Ye (2015)[7] as avaliações online são um fator importante na decisão de reserva dos hotéis, isto é, avaliações positivas aumentam a probabilidade de reserva e avaliações negativas diminuem. Essas avaliações podem ser encontradas em diversas plataformas, como TripAdvisor, Booking.com, Google, entre outras. Segundo Zhang et al. (2018)[8], os hotéis devem monitorizar e responder às avaliações online, oferecer incentivos para os clientes deixarem avaliações, pois para eles avaliações positivas têm um efeito maior do que as negativas, e o número de avaliações também influencia a decisão de reserva. Além desses fatores, outros fatores externos também poderão ter um impacto significativo nas reservas dos hotéis, tais como:

- **Períodos de férias:** Os períodos de férias podem ter um impacto significativo nas reservas de hotéis, especialmente em destinos turísticos populares. Durante os períodos de férias, a procura por hotéis aumenta de forma significativa, o que pode levar a um aumento nos preços. Os períodos de férias podem ser divididos em diferentes categorias, como alta temporada, baixa temporada e média temporada. Por exemplo, a alta temporada geralmente apresenta maior procura e preços mais elevados, enquanto a baixa temporada apresenta menor procura e preços mais baixos. Esses fatores devem ser considerados pelos hotéis ao desenvolver sua estratégia de preços e distribuição Kimes et al. (1998)[3]. A ocupação dos hotéis pode variar também de acordo com a época do ano, sendo que a procura é geralmente maior nos meses de verão, nas épocas festivas (como o Natal, por exemplo), e em feriados prolongados.

- **Eventos:** Eventos como grandes eventos culturais, festivais de música, exposições de arte, feiras gastronómicas, e eventos desportivos, podem influenciar as reservas de hotéis, pois atraem visitantes de fora da cidade ou do país. A procura durante esses eventos pode ser alta e os hotéis podem ajustar os seus preços conforme a procura e a importância do evento. Os hotéis podem trabalhar em parceria com os organizadores de eventos para oferecer pacotes especiais de alojamento.

- **Fatores meteorológicos**: As condições meteorológicas, como temperatura, precipitação e neve, podem influenciar as reservas de hotéis, especialmente em destinos de lazer. Por exemplo, em destinos de praia, a procura por hotéis pode aumentar durante os meses de verão, quando as temperaturas são mais quentes. Por outro lado, em destinos de neve, quando as temperaturas são mais frias, a procura pode ser maior para quem procura, por exemplo, desportos de inverno. Kilic et al (2005)[2].

- **Preços:** O preço dos hotéis pode ser um fator determinante na escolha

de um local de alojamento. Preços mais altos podem desencorajar alguns turistas de reservar quartos de hotel, enquanto preços mais baixos podem atrair uma maior quantidade de visitantes.

- **Qualidade das instalações e Avaliações:** A qualidade das instalações de um hotel, incluindo o nível de conforto, a limpeza, o atendimento ao cliente, a localização e as comodidades oferecidas, pode influenciar a decisão dos turistas em reservar um quarto no hotel, assim como as avaliações e recomendações efetuadas pelos clientes anteriores.

2.2 Glossário do Negócio

Termo	**Descrição**
e-GDS	Empresa de tecnologia especializada em soluções de distribuição global para gestão do comércio hoteleiro
Comércio hoteleiro	Conjunto de atividades relacionadas à gestão e operação de hotéis e outros estabelecimentos de hospedagem
Automatização de operações	Processo de substituição de atividades manuais por processos automatizados, visando aumentar eficiência e reduzir erros
Previsão de ocupação	Estimativa do número de quartos que estarão ocupados num determinado período
Reservas	Solicitações feitas pelos clientes para reservar quartos em um hotel
Booking.com	Plataforma online de reservas de hotéis e outros estabelecimentos de hospedagem

Afluência	Número de clientes que frequentam o hotel num determinado período
Cadeia hoteleira	Grupo de hotéis que operam sob uma mesma marca ou empresa
Gestão de reservas	Processo de receber, confirmar e organizar as reservas dos clientes
Integrações	Conexões entre diferentes plataformas ou sistemas que permitem a troca de informações
Ocupação	Percentagem de quartos ocupados relativamente ao número total de quartos disponíveis num determinado período
Parceiro tecnológico	Empresa que fornece soluções tecnológicas para outras empresas

3 Dados e-GDS

O *dataset* descrito nesta secção foi fornecido pela e-GDS e contém 4 tabelas com dados relacionados a reservas hoteleiras:

- Quartos reservados - Contém informações acerca das reservas em si
- Hotel - Contém os dados dos Hotéis
- Tipologias - Contém informações acerca dos quartos dos hotéis
- Facilities - Contém os dados das *facilities* dos hotéis

Cada um dos *datasets* é constituído por diferentes dados, sendo estes descritos nas tabelas seguintes. Todas as tabelas especificam o nome da coluna (Atributo), o tipo de dados, os valores permitidos (quando aplicável), se aceita valores nulos e a descrição da coluna.

Table 3.1: Dicionário de dados: Quartos Reservados

Atributo	Tipo	Valores Permitidos	Valores Nulos	Descrição
Hotel ID	int	>= 1	Não	Id do hotel
Reserve ID	int	>= 1	Não	Id da reserva
País	string	N/A	Não	País de orgiem da pessoa que reservou
Estado da reserva	string	Cancelado, Confirmado, CourtesyHold, Modificada, Não Registado, Pendente, Registado	Não	Estado em que a reserva se encontra
Room ID	int	>= 1	Não	ID do quarto reservado
Tipo de Quarto	string	N/A	Não	Tipo do quarto reservado
RatePlan	string	N/A	Não	
Data da reserva	string	YYYY-MM-DD HH:MM:SS.sss	Não	Data em que a reserva foi realizada
Data chegada	string	DD/MM/YYYY	Não	Data de chegada do cliente
Data de partida	string	DD/MM/YYYY	Não	Data de partida do cliente
Número de noites	int	>= 1	Não	Número de noites reservadas
Ocupação	int	>= 1	Não	Quantidade de pessoas da reserva
Adultos	int	>= 1	Não	Quantidade de adultos da reserva
Crianças	int	>= 0	Não	Quantidade de crianças da reserva
Bebés	int	>= 0	Não	Quantidade de bebés da reserva
Preço (€)	float	>= 0	Não	Custo da reserva em euros

Table 3.2: Dicionário de dados: Hotéis

Atributo	Tipo	Valores Permitidos	Valores Nulos	Descrição
Hotel ID	int	$>= 0$	Não	Id do hotel
Localização	string	N/A	Não	Localização do hotel
Estrelas	int	0, 1, 2, 3, 4, 5	Não	Estrelas do hotel
Idade Máxima de Crianças	int	0, 2, 3, 5, 6, 8, 10, 11, 12, 13, 15, 16, 17, 18, 36	Não	Idade limite em que uma pessoa é considerada criança em anos
Idade Máxima de Bebés	int	0, 1, 2, 3, 4, 11, 12, 23, 24, 36, 48, 60, 168	Não	Idade limite em que uma pessoa é considerada bebé em meses
Hora máxima de check-in	string	Hora com minutos e segundos	Não	Hora máxima em que as pessoas podem dar check in
Quantidade de quartos	int	$>= 1$	Não	Quantidade de quartos que o hotel possui

Table 3.3: Dicionário de dados: Tipologias

Atributo	Tipo	Valores Permitidos	Valores Nulos	Descrição
Hotel ID	int	$>= 0$	Não	Id do hotel
Room ID	int	$>= 0$	Não	Id do quarto
Tipo de quarto	string	Qualquer string	Não	Tipo do quarto
Quantidade	int	$>= 0$	Não	Quantidade de quartos existentes
Capacidade máxima	int	$>= 0$	Não	Capacidade máxima em termos de ocupantes
Capacidade mínima	int	$>= 0$	Não	Capacidade mínima em termos de ocupantes
Capacidade máxima de adultos	int	$>= 0$	Não	Capacidade máxima de adultos
Capacidade mínima de adultos	int	$>= 0$	Não	Capacidade mínima de adultos
Capacidade máxima de crianças	int	$>= 0$	Não	Capacidade máxima de crianças
Capacidade mínima de crianças	int	$>= 0$	Não	Capacidade mínima de crianças
Capacidade máxima de bebés	int	$>= 0$	Não	Capacidade máxima de bebés
Capacidade máxima de camas extra	int	$>= 0$	Não	Capacidade máxima de camas extra
Capacidade máxima de camas extra (crianças)	int	$>= 0$	Não	Capacidade máxima de camas extra para crianças
Capacidade máxima de berços extra	int	$>= 0$	Não	Capacidade máxima de berços extra

Table 3.4: Dicionário de dados: Facilities

Atributo	Tipo	Valores Permitidos	Valores Nulos	Descrição
Hotel ID	int	$>= 0$	Não	Id do hotel
Facility ID	int	$>= 0$	Não	Id da facility
Nome	string	Qualquer string	Não	Nome da facility

4 Fontes de dados externas

4.1 Feriados

dataset criado através da utilização do serviço sapo [6]. Este serviço fornece os feriados existentes para um determinado ano, desta forma foi criado um ficheiro em formato csv com todos os feriados desde 2022 até 2023. Este *dataset* será relevante uma vez que a afluência dos hotéis aumenta nos dias de feriados ou fins de semana.

Table 4.1: Dicionário de dados: Feriados

Atributo	Tipo	Valores Permitidos	Valores Nulos	Descrição
Date	date	dd/MM/aaaa	Não	Data
day	int	1 - 31	Não	Dia do mês
dayOfWeek	int	1 - 7	Não	Dia da semana
month	int	1 - 12	Não	Mês
trimester	int	1, 2, 3 ou 4	Não	Trimestre
year	int	2011 - 2023	Não	Ano
isHoliday	boolean	0 ou 1	Não	Se é feriado

4.2 Meteorologia

Este *dataset* foi criado com recurso ao website Meteostat [5] que contêm uma enorme quantidade de dados sobre a meteorologia global. Desta forma, procuramos pela cidade que corresponde à localização mais próxima do hotel e nas datas compreendidas entre *01-01-2022* e *23-04-2023* para que possam coincidir com as datas das reservas. Este *dataset* é importante, uma vez que a afluência dos hotéis poderá aumentar nos dias de calor ou de maior temperatura e diminuir nos dias de temperatura mais baixas.

Table 4.2: Dicionário de dados: Meteorologia

Atributo	Tipo	Valores Permitidos	Valores Nulos	Descrição
date	date	aaaa-mm-dd	Não	Data
tavg	float	Qualquer valor	Sim	Temperatura média
tmin	float	Qualquer valor	Sim	Temperatura mínima
tmax	float	Qualquer valor	Sim	Temperatura máxima
prcp	float	>= 0	Sim	Total de precipitação
wdir	int	>= 0 e <= 360	Sim	Direção do vento
wspd	float	>= 0	Sim	Velocidade do vento
wpgt	float	>= 0	Sim	Pico de rajada
pres	float	>= 0	Sim	Pressão do ar
city	string	Qualquer string	Não	Cidade correspondente à localização do hotel

4.2.1 Alterações e Limpeza

- Foram eliminadas as linhas que apenas continham a data e a cidade e não continham dados meteorológicos.
- Foi adicionado uma nova coluna *city* ao *dataset* para permitir integrar os dados meteorológicos com as reservas.

- Foi eliminado a coluna *snow* pois a profundidade da neve não é relevante para o contexto do problema.

4.3 Eventos

dataset de eventos que foi elaborado com pesquisa e com recurso também do ChatGPT [1] que indicou eventos relevantes e úteis para o problema. Os eventos serão talvez se não o *dataset* com maior relevância e importância na afluência dos hotéis. Isto porque, um grande evento pode levar ao alojamento de milhares de pessoas naquela localização.

Table 4.3: Dicionário de dados: Eventos

Atributo	**Tipo**	**Valores Permitidos**	**Valores Nulos**	**Descrição**
Location	string	Qualquer string	Não	Cidade/Localização do evento
Event	string	Qualquer string	Não	Evento
Start_date	date	dd/mm/aaaa	Não	Data de ínicio do evento
End_date	date	dd/mm/aaaa	Não	Data de fim do evento

5 Integração dos dados

Com o objetivo de obter um único *dataset* final, foi seguido um processo de integração de todos os *datasets*. Destes *datasets*, 4 foram fornecidos pela e-GDS, dados de reservas, de hotéis, de tipologias e facilities. Decidimos descartar o *dataset* das facilities uma vez que apenas foi dado o nome das facilites por hotel e a grande maioria, se não todos, os nomes são diferentes uns dos outros, impossibilitando a sua compreensão e tratamento. Adicionalmente foram obtidos 3 *datasets* externos de fontes públicas sobre feriados, meteorologia e eventos.

Primeiramente começamos por juntar numa só tabela os dados da e-GDS. Exportamos as tabelas fornecidas em excel para um ficheiro csv e com recurso ao spark lemos estes ficheiros e a partir deles criamos 4 *datasets* em memória. [1] Seguidamente à operação de leitura, foram feitas algumas alterações, sendo estas a renomeação de colunas para remover caracteres especiais e a transformação dessas mesmas colunas no seu tipo de dados correto. De forma similar, este processo de leitura foi aplicado aos *datasets* externos.

Ainda durante a leitura dos dados dos hotéis foi adicionada uma nova coluna (area_localizacao), baseada na localização do hotel, que permite a integração com o *dataset* da meteorologia. Esta coluna mapeia a localização do hotel para a localização para o qual o *dataset* de meteorologia têm dados.

A segunda etapa consistiu em juntar estes *datasets* com base nas colunas idênticas [2]. Juntou-se assim o *dataset* das reservas com o *dataset* das tipolo-

[1] Em anexo "LeituraDatasets.zpln"
[2] Em anexo "DatasetsJoin.zpln"

gias com base nas colunas Hotel ID e Room ID. Inicialmente no *dataset* de reservas existiam $25,090$ entradas e depois desta junção, como nem todos os hotéis e quartos tinham a tipologia definida, o número de entradas reduziu para $25,081$.

De seguida juntamos este *dataset* com o *dataset* que contém a informação dos hotéis com base na coluna Hotel ID. Com esta junção o número de entradas manteve-se. Podemos perceber que todos os hotéis referenciados nas reservas existem na tabela dos hotéis. Desta forma, finalizamos a integração dos dados fornecidos pela e-GDS.

Faltando juntar os dados das fontes externas, começamos por integrar os feriados. Decidimos juntar as reservas com os feriados que ocorressem durante a estadia no hotel. Para tal, fizemos um left join da tabela resultante do passo anterior com a dos feriados onde a data de partida seja maior ou igual à data do feriado e a data de chegada seja menor ou igual à data do feriado. Com esta junção o número de linhas aumentou para $25,451$ uma vez que existem reservas com mais que um feriado entre a data de chegada e de partida. Esta integração permitirá saber quais as reservas agendadas durante feriados.

A seguir foi feito o join com a tabela de eventos, para tal a coluna criada na tabela dos hotéis (area_localizacao) foi usada com a coluna 'Localização' do *dataset* dos eventos. Com esta integração o número de entradas aumentou para $25,481$ uma vez que existem reservas com mais que um evento próximo durante a estadia. Esta integração permitirá saber que eventos existiam na altura da reserva.

Por fim foi feito o join com o *dataset* da meteorologia com base nas colunas area_localizacao da tabela anterior e city da tabela da meteorologia e também com base nas datas de partida e chegada das reservas. Com esta integração o *dataset* passou a ter $78,020$ entradas, criando para cada reserva o mesmo número de entradas quanto o número de dias. Esta integração adiciona dados meteorológicos às reservas.

Com a integração completa, o próximo passo será agrupar as linhas conforme a expansão provocada pela integração dos *datasets* externos. Por exemplo, agrupar as entradas pelo id da reserva e calcular a média de temperatura, o número de eventos e feriados. No fim desse processo de tratamento será expectável obter o mesmo número de linhas do *dataset* original.

6 Análise dos dados

6.1 E-GDS

6.1.1 Análise Inicial

De forma a analisar os dados eles foram lidos em Spark, como mencionado na secção anterior e algumas modificações foram feitas. Nomes de colunas foram alterados e os tipos de dados foram transformados no tipo correto.

Table 6.1: Descrição da tabela Hotel

summary	hotel_ID	localizacao	estrelas	idade_max_criancas	idade_max_bebes	qtd_quartos	area_localizacao
count	145	145	145	145	145	145	145
mean	N/A	N/A	1,79	7,32	20,02	30,02	N/A
stddev	N/A	N/A	1,78	6,15	20,01	33,80	N/A
min	20	N/A	0	0	0	1	N/A
max	561	N/A	5	36	168	192	N/A

O conjunto de dados apresentado na tabela 6.1 é composto por informações de 145 hotéis e 7 colunas distintas. As colunas incluem o ID do hotel, sua localização, número de estrelas, idade máxima para crianças e bebés, número de quartos e área de localização. A média de estrelas dos hotéis é de 1,79, indicando que a maioria dos hotéis possui menos de 2 estrelas. Não foram identificados valores nulos na tabela. Porém, é importante ressaltar que foi identificado um valor errado na idade máxima para crianças, que foi de 36 anos. Além disso, alguns hotéis foram registados com localizações anormais, como ”.” e ”teste”. Ao verificar se esses hotéis possuíam reservas, constatou-se

que 57 hotéis não contavam com nenhuma reserva registada.

Table 6.2: Descrição da tabela Tipologia 1/2

summary	hotel_ID	room_ID	tipo_quarto	quantidade	capacidade_maxima	capacidade_minima	capacidade_max_adultos
count	741	741	741,00	741	741	741	741
mean	N/A	N/A	N/A	5,22	3	1,09	2,75
stddev	N/A	N/A	N/A	8,71	1,43	0,58	1,34
min	20	81	N/A	0	1	1	1
max	561	2905	N/A	90	10	10	10

Table 6.3: Descrição da tabela Tipologias 2/2

summary	capacidade_max_criancas	Capacidade_mx _bebés	capacidade_max_camas_extra	capacidade_mx_camas_criancas	capacidade_max_bercos
count	741	741	741,00	741	741
mean	1,08	0,59	0,14	0,15	0,361
stddev	1,38	0,64	0,38	0,4	0,55
min	0	0	0,00	0	0
max	8	3	2,00	2	2

Este dataset que se encontra na tabela 6.2 e 6.3 refere-se à tabela tipologia e contém como principais atributos as colunas tipo de quarto, quantidade, capacidade máxima e mínima, e capacidade máxima de adultos e crianças. Existem 741 tipos de quartos diferentes. A capacidade máxima média é de 3 pessoas, enquanto a capacidade mínima média é de 1 pessoa. Os valores mínimo e máximo variam em função das características de cada tipo de quarto, com um mínimo de 0 e um máximo de 90 para a quantidade de quartos. A capacidade máxima varia de 1 a 10 pessoas, enquanto a capacidade máxima de adultos varia de 1 a 10 e a capacidade máxima de crianças varia de 0 a 8. De notar que a tabela não contém valores nulos. Numa primeira observação verificamos que poderão existir campos irrelevantes, sendo eles os seguintes:

- "Capacidade mínima": irrelevante mais de 90% tem o valor de "1";
- "Capacidade mínima de adultos": irrelevante mais de 90% o valor de "1"
- "Capacidade mínima de crianças": irrelevante mais de 95% tem o valor de "0"

- "Capacidade máxima de camas extra": irrelevante, é um valor subintendido na capacidade máxima de adultos, bebés e crianças;
- "Capacidade máxima de camas extra (crianças)": irrelevante, é um valor subintendido na capacidade máxima de adultos, bebés e crianças;
- "Capacidade máxima de berços extra": irrelevante, é um valor subintendido na capacidade máxima de adultos, bebés e crianças;

Table 6.4: Descrição da tabela Quartos Reservados 1/2

summary	hotel_ID	Reserve_ID	pais	estado_reserva	room_ID	tipo_quarto	rate_plan	num_noites
count	25105	25105	25105	25105	25105	25105	25105	25105
mean	N/A	N/A	N/A	N/A	1944,56	N/A	N/A	2,28
stddev	N/A	N/A	N/A	N/A	621,15	N/A	N/A	4,68
min	20	1418210	N/A	N/A	81	N/A	N/A	1
max	561	1723815	N/A	N/A	2897	N/A	N/A	481

Table 6.5: Descrição da tabela Quartos Reservados 2/2

summary	ocupacao	adultos	criancas	bebes	preco_euros	data_reserva	data_chegada	data_partida
count	25105	25105	25105	25105	25105	730	730	730
mean	1,87	2,04	0,06	1,87	243,33	N/A	N/A	N/A
stddev	0,66	1,02	0,30	0,66	373,54	N/A	N/A	N/A
min	1	1	0	1	0	01/01/2022	01/01/2022	02/01/2022
max	9	30	4	9	20683	31/12/2023	05/04/2024	31/05/2024

O conjunto de dados apresentados na tabela 6.4 e 6.5 é o nosso dataset principal. Este contém informações sobre reservas de hotéis, incluindo detalhes como país, estado da reserva, tipo de quarto, plano de tarifas, número de noites, ocupação, número de adultos, crianças e bebés, preço em euros, e datas de reserva, chegada e partida. Ao analisar o conjunto de dados, podemos ver que há 25.105 entradas e 13 atributos. A média de noites reservadas é de 2,28, com um desvio padrão de 4,68, indicando uma grande variação nos períodos

de estadia. Além disso, a média de ocupação é de 1,87, o que significa que, em média, cada quarto é ocupado entre uma a duas pessoas. No entanto, não conseguimos perceber se essa ocupação está correta, uma vez que não é coerente com a soma de adultos e crianças. Vamos optar por assumir a ocupação como dado correto. Verifica-se que os dados referentes a bebés deverá estar errado por quase todas as reservas terem incluídos bebés, já com as crianças essa situação não se verifica. Relativamente aos preços em euros, a média é de 243,33 com um desvio padrão de 373,54, mostrando que há uma grande variação de preços nas reservas de hotéis. O número máximo de noites reservadas é de 481, o que pode indicar que algumas pessoas reservam hotéis por longos períodos de tempo, como para estadias prolongadas ou negócios, no entanto, como a média de reservas é 2,28 podemos verificar que o número 481 é um outlier. As datas de reserva variam de 01/01/2022 a 31/12/2023, com datas de chegada e partida que vão até 05/04/2024 e 31/05/2024, respetivamente. Isso sugere que as reservas foram feitas com bastante antecedência e que há uma ampla janela de tempo para reservar um quarto de hotel. Como análise complementar fizemos uma verificação direcionada para os seguintes campos:

1. Datas: Para verificar se as datas do dataset se encontravam coerentes, foi analisado se as datas de saída do hotel eram sempre superiores às datas de entrada e se as datas de reserva eram inferiores às datas de chegada. Também foi calculado manualmente o número de noites atráves da data de chegada e de saída e todos os valores se encontravam corretos.

2. RatePlan: A partir dos dados apresentados na tabela 6.6, podemos concluir que os “Rateplans” mais frequentes nas reservas são ”MR - Main Rate” e ”Normal”, ambos com mais de 2400 reservas. Esses resultados sugerem que existe uma grande quantidade de reservas feitas com os mesmos rate plans, o que pode ter implicações na estratégia de preços e de previsão de reservas com base no rate plan

Table 6.6: Tabela com o 10 RatePlan mais Relevantes

Rate Plan	**Contagem**
MR - Main Rate	2545
Normal	2417
BAR	1827
WEB (Best Available Rate)	1707
Bar	1693
WebSite	1681
Main Rate BB	1057
Standard	557
(WEB) Best Available Rate	556
Não Reembolsável	401

3. Preço:

hotel ID	**preco_medio_por_noite**
284	51.18
328	52.85
522	56.74
560	57.54
283	58.35

(a) Hoteís baratos

hotel_ID	**preco_medio_por_noite**
358	263.39
445	252.59
513	250.0
443	237.41
225	235.01

(b) Hoteís caros

Figure 6.1: Comparação entre preço por noite em hoteís caros e baratos

O preço é um fator importante a ser considerado ao reservar um hotel, pois pode impactar diretamente o número de reservas feitas. Com isso em mente, realizamos uma agregação de preços divididos pelo número de noites em cada hotel, permitindo-nos obter o preço médio por noite

durante o período de estudo. As tabelas 6.1a e 6.1b apresentam os hotéis com os preços médios mais baixos e mais altos por noite, respetivamente. É possível observar que o hotel com o preço mais baixo é o de ID 284, com um valor médio de 51,18 euros por noite. Em contrapartida, o hotel com o preço mais alto é o de ID 358, com um valor médio de 263,39 euros por noite.

4. País: Para termos uma compreensão mais abrangente das reservas de

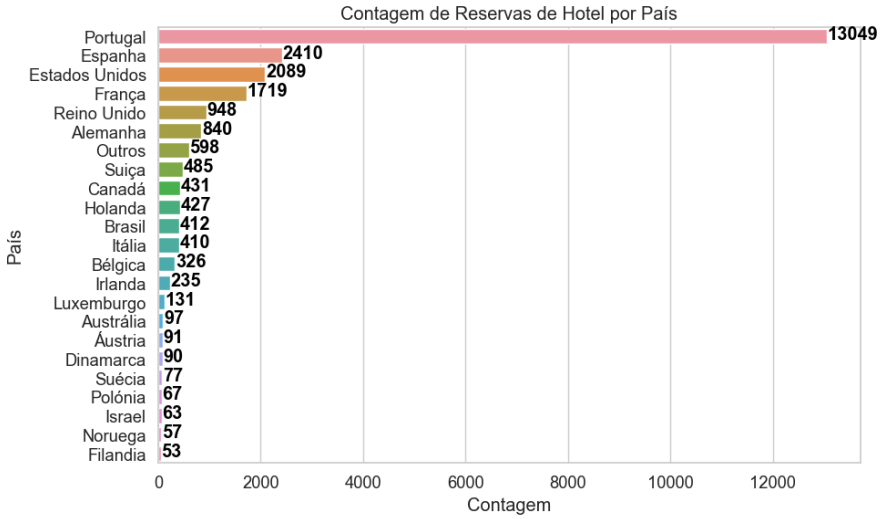


Figure 6.2: Reservas Por País: Países com mais reservas

hotéis e das nacionalidades dos hóspedes, a agregação dos dados por país foi uma etapa essencial. Inicialmente, o conjunto de dados possuía informações de várias nacionalidades, mas para melhor visualização dos resultados, foi necessário agrupá-las por país e ordená-las por número de reservas. A contagem de reservas por país foi realizada e utilizada para criar gráficos que destacam os países com maior número de reservas. Para melhorar a visualização dos dados como mostra a figura 6.2, os países com menos de 50 reservas foram agrupados em uma categoria denominada ”Outros”, permitindo uma melhor compreensão dos dados e identificação dos países com maior e menor número de reservas. Com

isso, descobrimos que Portugal é o país com maior número de reservas de hotéis, com 13049 reservas, e a Bulgária é o país com o menor número de reservas, com apenas 53 reservas. O agrupamento dos países menos frequentes numa única categoria "Outros", resultou em 598 reservas, com 43 países com apenas 1 ou 2 reservas. Esses países podem ser considerados outliers em futuras análises de previsão de reservas, mas neste momento com essa agregação outros, chegamos à conclusão que não vamos remover qualquer país sem uma análise pais profunda primeiro. Em resumo, a agregação por país na nossa perspetiva é uma etapa crucial para a compreensão dos dados e identificação dos países com maior e menor número de reservas, permitindo a tomada de decisões mais informadas relativamente a uma futura previsão de reservas.

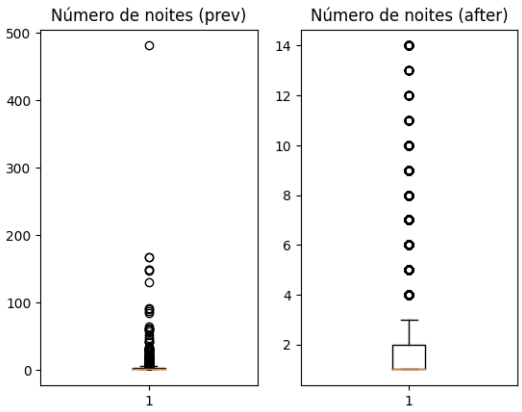
6.1.2 Análise de Outliers

A remoção de outliers é uma técnica comumente utilizada para eliminar valores discrepantes que podem distorcer a análise estatística dos dados e afetar a precisão das conclusões obtidas a partir deles. A existência destes valores pode também influenciar negativamente a qualidade de um modelo de machine learning.

Número de Noites e Preço

Identificado que a coluna "Número de Noites" continha valores muito divergentes, que eram notavelmente outliers. Para verificar os outliers no número de noites, utilizou-se o Three Sigma, uma técnica estatística que identifica valores discrepantes, para isso, aplicou-se a seguinte fórmula:

- (Limite Superior = média + 3 * desvio padrão)
- Média Número noites: 2,28;
- Desvio padrão número de Noites: 4,68



(a) Análise Outliers: Número de noites

(b) Análise Outliers: Preço

- Limite superior = 2.28 + 3 * 4.68
- Limite superior = 14,32

Este resultado foi arredondado para 14. No limite inferior, consideramos o valor 0, uma vez que, não existem reservas negativas. Assim, aplicando o filtro ao número de noites, temos o total de 115 reservas. Para efetuar a análise dos outliers foi criada uma biblioteca em python de suporte [1].

Como mostrado nas figuras 6.3a e 6.3b. Para o número de noites existem 115 reservas com um valor superior a 14, que por sua vez dá para verificar que esse número de noites anormal afeta o preço. Com esta análise, verifica-se que existem 7 reservas com valores superiores a 5000 Euros.

Número de Reservas

Com base na análise das reservas por hotel, podemos observar uma grande variação nos números de reservas, com alguns hotéis apresentando um número significativamente menor de reservas do que outros. Para garantir uma análise mais robusta, decidimos considerar como outliers os hotéis com um número de reservas inferior a 13, com base no critério de que esses hotéis apresentam

[1] Em anexo "Analises.py"

uma frequência muito baixa. Dessa forma, identificamos cinco hotéis como outliers, o que resultou na remoção de 20 reservas da análise. A tabela 6.8 apresenta os hotéis identificados como outliers, juntamente com suas localizações e o número de reservas. A escolha do número 13 foi feita com base na análise

Quartil	**Valor**
0.00	1.00
0.25	45.75
0.50	130.00
0.75	229.00
1.00	3325.00

Table 6.7: Tabela de Quartis

dos quartis como mostra a tabela 6.7 de quantidade de reservas por hotel, verificamos que o primeiro quartil tem um valor de 45,75 reservas por hotel. No entanto, consideramos que remover todos os hotéis abaixo desse valor teria um impacto significativo na amostra, resultando na eliminação de um número considerável de hotéis. Dessa forma, escolhemos um valor intermédio de 13 reservas. Vale ressaltar que a escolha desse valor foi feita com base numa avaliação cuidadosa das implicações práticas dessa decisão, procurando minimizar o impacto na amostra e preservar a representatividade dos dados.

Table 6.8: Hoteis Com menos reservas

hotel_ID	**localizacao**	**count**
302	Tomar	5
491	Vilamoura	5
390	Alijó	5
442	Funchal	4
513	Tavira	1

6.2 Feriados

O *dataset* Feriados que se encontra na tabela 6.9 apresenta informações sobre as datas presentes no conjunto de dados, tais como dia, dia da semana, mês, trimestre, ano, indicação de feriado e nome da semana em português. Em resumo, a tabela fornece informações sobre as datas presentes no conjunto de dados, com destaque para os dias de fim de semana e feriados. A informação deste *dataset* poderá ser muito útil para perceber se existe mais reservas durante os períodos com feriados ou fim de semana.

Table 6.9: Descrição da tabela Feriados

summary	day	dayOfWeek	month	trimester	year	is_holiday	portugueseWeekName	date
count	730	730	730	730	730	730	730	730
mean	15,7	4	6,53	2,51	2022,50	0,04	N/A	N/A
stddev	8,8	2	3,45	1,12	0,50	0,19	N/A	N/A
min	1	1	1	1	2022	0	N/A	01/01/2022
max	31	7	12	4	2023	1	N/A	31/12/2023

6.3 Meteorologia

O *dataset* Meteorologia contêm dados meteorológicos de 28 localizações diferentes compreendidos entre as datas *2022-01-01* e *2023-04-23*. Estas localizações correspondem com as zonas onde os hóteis se encontram. Como se segue na tabela 6.10, este *dataset* fornece-nos informações como temperatura média, máxima, mínima, precipitação total daquele dia, informações do vento e a cidade onde os valores foram registados. Este dataset apresenta também poucos valores nulos e permite-nos retirar informações como:

- Maior temperatura registada: 44.3
- Menor temperatura registada: -4.7

- Temperatura média de todos os registos: 15.59

Table 6.10: Descrição da tabela Meteorologia

summary	tavg	tmin	tmax	prcp	wdir	wspd	wpgt	pres	city	date	
count	13293	13285	13289	9935	13275	13275	8352	13275	13293		13275
mean	15,59	11,65	20,13	2,57	193,56	11,67	31,49	1019,3	N/A	N/A	
stddev	5,13	5,19	6,01	6,98	112,44	6,09	9,95	6,4	N/A	N/A	
min	-0,3	-4,7	3,1	0	0	1,4	7,4	992,9	N/A	01/01/2022	
max	35,1	28,6	44,3	118	360	49,1	87	1040,7	N/A	23/04/2023	

6.4 Eventos

Este *dataset* contêm eventos relevantes compreendidos entre as datas de *01/06/2022* e *05/11/2023* e a respetiva localização para que seja possível relacionar com as reservas e meteorologia. Da descricão que se segue 6.11 apenas conseguimos retirar que contamos com 59 eventos diferentes compreendidos nas datas mencionadas acima. Neste *dataset* não existem valroes nulos.

Ï

Table 6.11: Descrição da tabela Eventos

summary	Location	Event	start_Date	end_date
count	59	59	59	59
mean	N/A	N/A	N/A	N/A
stddev	N/A	N/A	N/A	N/A
min	N/A	N/A	01/06/2022	11/06/2022
max	N/A	N/A	05/11/2023	05/11/2023

7 Sugestões de Tratamento de dados

Neste capítulo, apresentamos um resumo das sugestões de tratamento de dados para cada um dos conjuntos de dados analisados nos capítulos anteriores: Hotél, Tipologias, Facilities e Quartos Reservados. Foram realizadas diversas ações, tais como remoção de colunas irrelevantes, remoção de outliers e adição de novas colunas. O objetivo é obter um conjunto de dados mais limpo e coerente para uma melhor análise do dataset principal

7.1 Dataset Hotel

- Remoção de 57 linhas de hotéis para os quais não existem reservas

7.2 Dataset Tipologias

- Remoção da coluna “Capacidade mínima”: irrelevante mais de 90% tem o valor de “1”;
- Remoção da coluna “Capacidade mínima de adultos”: irrelevante mais de 90% tem o valor de “1”
- Remoção da coluna “Capacidade mínima de crianças”: irrelevante mais de 95% tem o valor de “0”
- Remoção da coluna “Capacidade máxima de camas extra”: irrelevante, é um valor subintendido na capacidade máxima de adultos, bebés e crianças;

- Remoção da coluna “Capacidade máxima de camas extra (crianças)”: irrelevante, é um valor subintendido na capacidade máxima de adultos, bebés e crianças;
- Remoção da coluna “Capacidade máxima de berços extra”: irrelevante, é um valor subintendido na capacidade máxima de adultos, bebés e crianças;

7.3 Dataset Facilities

Não será feita nenhuma limpeza uma vez que não se consegui identificar como poderão ser usados os dados. Alguns hotéis têm muitas facilites, outros poucas. Além disso, existe uma grande quantidade de variedade dos dados, sendo que são poucos os hotéis que possuem facilites similares.

7.4 Dataset Quartos Reservados

- Remoção de 5 linhas com estado de reserva ”CourtesyHold”
- Remoção de 115 linhas onde o número de noites é superior a 14 (outliers)
- Remoção de cinco hotéis, que equivale a 15 linhas de reservas, onde o número de reservas por hotel é inferior a 13 (outliers)
- Remoção de 5 linhas onde o preço é maior que 5000 €
- Adição da coluna ”preco_por_noite”
- Adição de coluna ”area_localizacao” para permitir fazer a integração dataset meteorologia e eventos;

7.5 Tratamentos finais

- Agregação do dataset resultante pelo ID da reserva com o seguinte tratamento:

- Remoção de todas as colunas relativas aos feriados, exceto a coluna is_holiday.
- Remoção das colunas dos eventos e adição da coluna número de eventos que retrata a quantidade de eventos durante a reserva.
- Quanto à meteorologia, iremos acrescentar colunas relativas à temperatura mínima da reserva, média e máxima. Para a precipitação também teremos a mínima, média e máxima.

Bibliography

[1] Chat.openai.com. https://chat.openai.com/, 2023.

[2] Huseyin Kilic and Fevzi Okumus. The effect of service quality on customer loyalty within the context of ski resorts. *Journal of Hospitality & Leisure Marketing*, 12(3):23–43, 2005.

[3] Sheryl E Kimes, Jochen Wirtz, and Betsy M Noone. Restaurants and hotels: Strategies for the hospitality industry. *Cornell Hotel and Restaurant Administration Quarterly*, 39(3):60–68, 1998.

[4] Jae-Hyeon Lee and Byong-Hun Jeon. Determinants of customer satisfaction and loyalty in the korean hotel industry. *Journal of Hospitality & Tourism Research*, 36(3):389–417, 2012.

[5] Meteostat.net. https://meteostat.net/en/, 2023.

[6] Sapo.pt. https://services.sapo.pt/Holiday/GetNationalHolidays?year=2023, 2023.

[7] Fang Xu and Qiang Ye. Effects of online reviews on hotel booking intention: The moderating role of hotel price. *Journal of Hospitality & Tourism Research*, 39(1):3–23, 2015.

[8] Xinyuan Zhang, Lijuan Huang, and Xin Zhao. The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management*, 67:287–308, 2018.