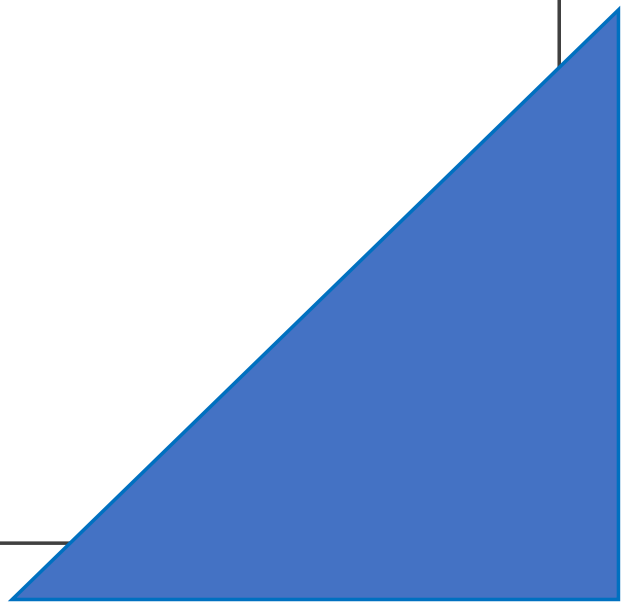
TEAD - TP

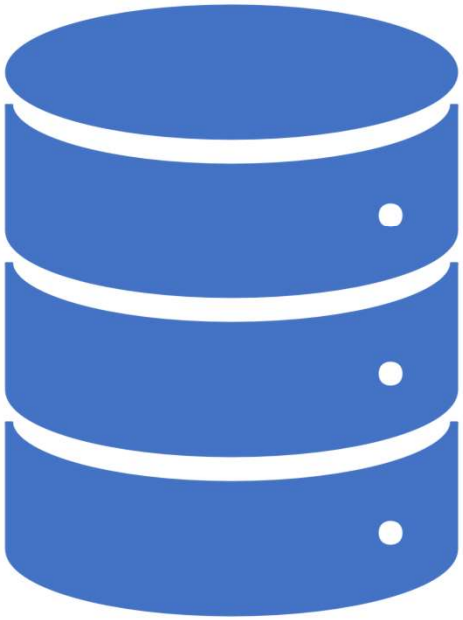
João Bragança – 8190555

José Fernandes – 8190239

Pedro Afonso – 8090457



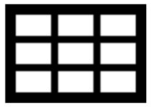
Introdução do problema



Data Set Original

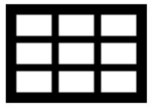
Data Set Original

Quartos Reservados



Hotel

Tipologias



Facilities

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

Atributo	Tipo	Valores Permitidos	Valores Nulos	Descrição
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

Hotel

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	Não	Idade limite em que uma pessoa é considerada criança em anos
Idade Máxima de Bebés	int	>= 0	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

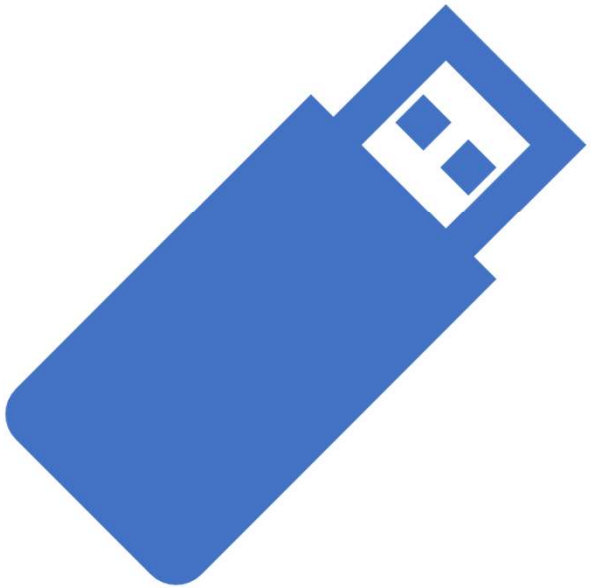
Tipologias

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

Atributo	Tipo	Valores Permitidos	Valores Nulos	Descrição
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

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



Fontes externas

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

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

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

Análise inicial dos dados

Quartos reservados

- Interpretação errada da coluna ocupação.
- Reservas apenas a partir de 2022 (o que torna difícil a previsão)
- *Rate plans* iguais com nomes diferentes

Quartos reservados

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

Quartos reservados

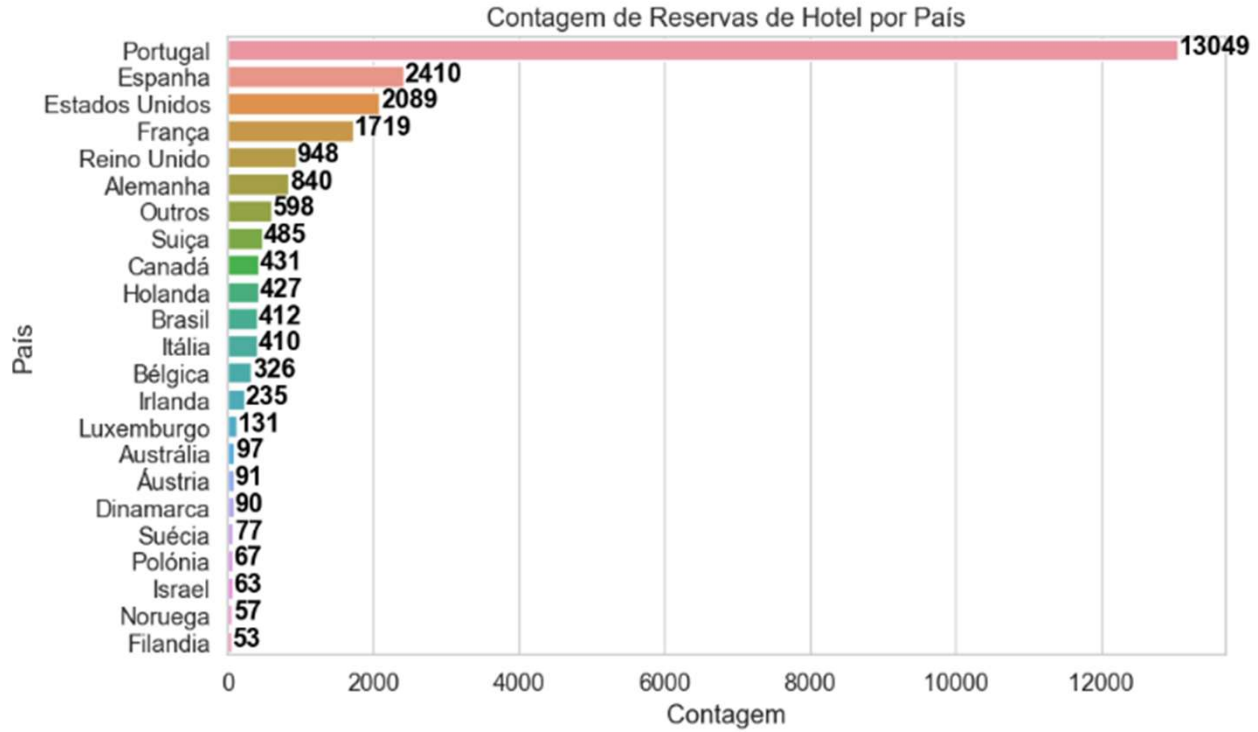
hotel ID	preco_medio_por_noite
284	51.18
328	52.85
522	56.74
560	57.54
283	58.35

(a) Hoteis baratos

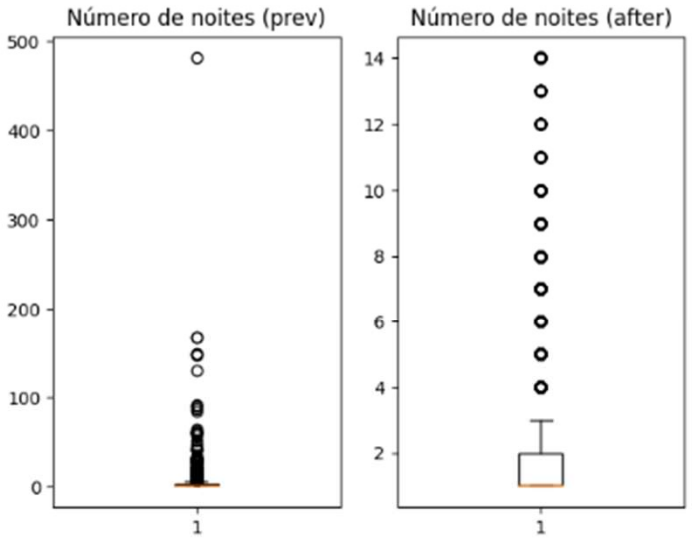
hotel_ID	preco_medio_por_noite
358	263.39
445	252.59
513	250.0
443	237.41
225	235.01

(b) Hoteis caros

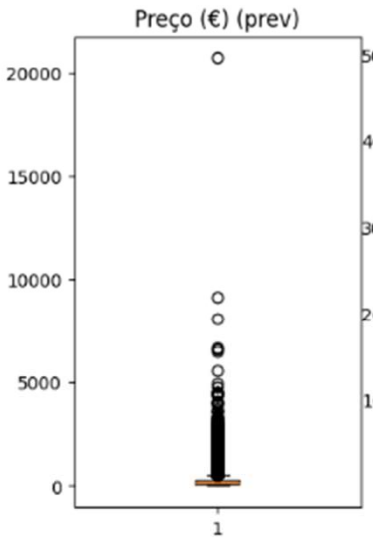
Quartos reservados

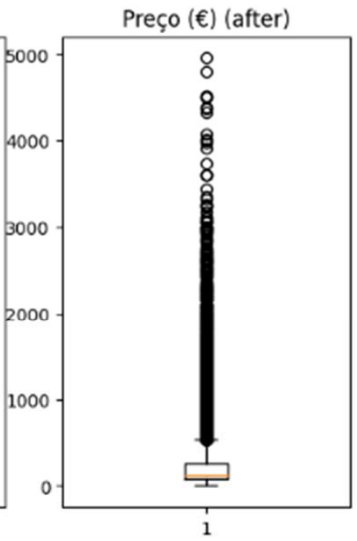


Quartos reservados



(a) Análise Outliers: Número de noites





(b) Análise Outliers: Preço

Quartos reservados

Tabela 5.8: Hoteis Com menos reservas

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

Hotel

- Alguns valores com falhas:
  - 36 como idade máxima de crianças
  - Localizações “teste” e “.”
  - 57 hotéis sem reservas

Tipologias

- Colunas possivelmente irrelevantes:
  - Capacidade mínima: mais de 90% com o valor “1”
  - Capacidade mínima de adultos: mais de 90% com o valor “1”
  - Capacidade mínima de crianças: mais de 95% com o valor “0”

Facilities

- Dados muito divergentes:
  - Alguns hotéis com muitas facilities, outros sem nenhuma
  - Facilities com importâncias muito diferentes (spa, free wi fi)

Tratamento dos dados

Data Set Hotel

- Remoção de 57 linhas de hotéis para os quais não existem reservas
- Adição de coluna "area_localizacao" para permitir fazer a integração com os data sets meteorologia e eventos

Data Set Tipologias

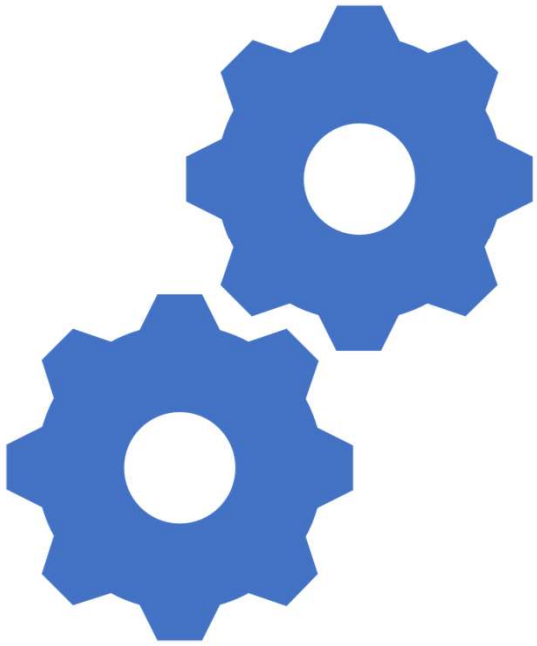
- Remoção da coluna “Capacidade mínima”: irrelevante mais de 90% tem o valor de “1”
- Remoção da coluna “Capacidade mínima de adultos”: irrelevante mais de 90% tem o valor de “1”
- Remoção da coluna “Capacidade mínima de crianças”: irrelevante mais de 95% tem o valor de “0”

Data Set Facilities

- Não foi feita nenhuma limpeza uma vez que não se conseguiu identificar como poderão ser usados os dados.

Data Set Quartos Reservados

- Remoção de 4 linhas com estado de reserva *CourtesyHold*.
- Remoção de 115 linhas onde o número de noites é superior a 14.
- Remoção de cinco hotéis, onde o número de reservas por hotel é inferior a 13
- Remoção de 1 linha onde o preço por noite por ocupação é maior ou igual a 1000€
- Remoção de 10 linhas duplicadas
- Adicionada uma coluna chamada *dif_data_chegada_data_reserva*
- Normalização da coluna *rate_plan.* 230 -> 44
- Normalização da coluna *Tipo de quarto* 344 -> 82



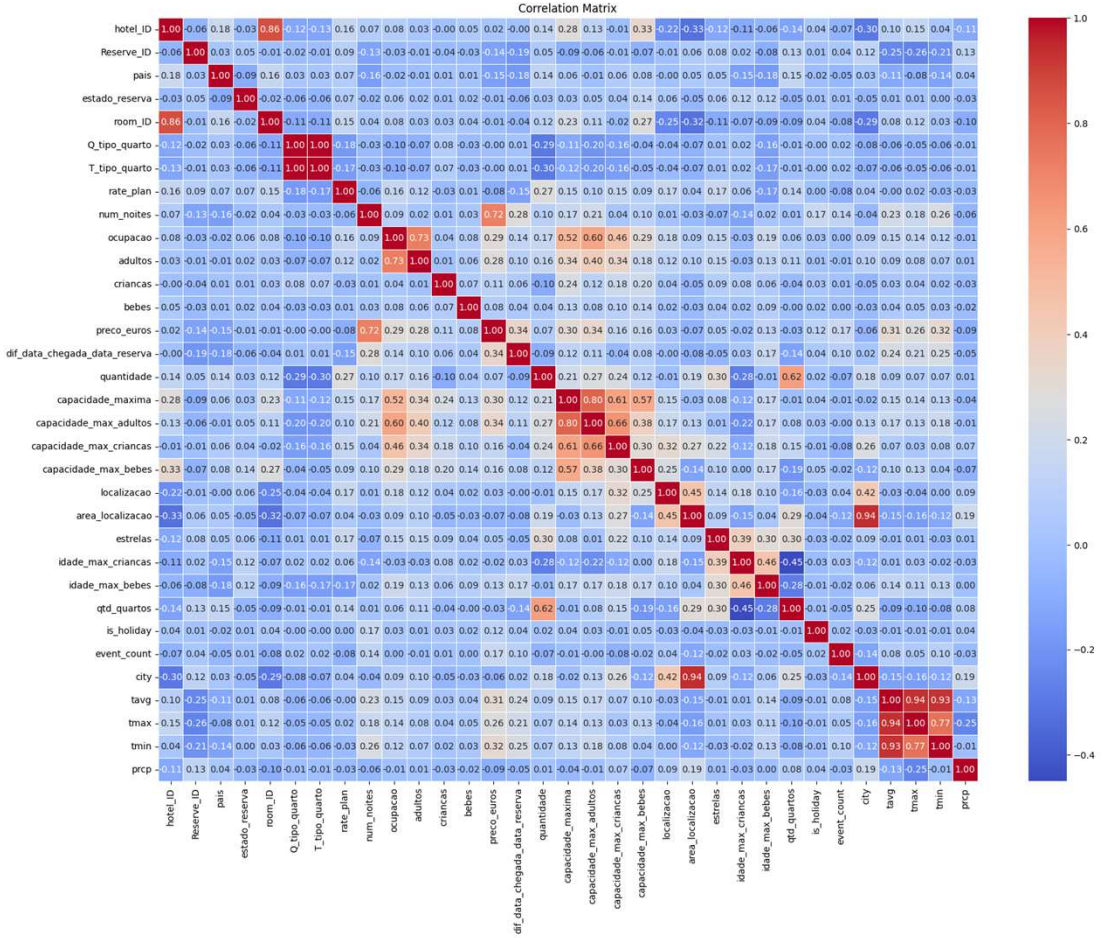
Integração dos dados

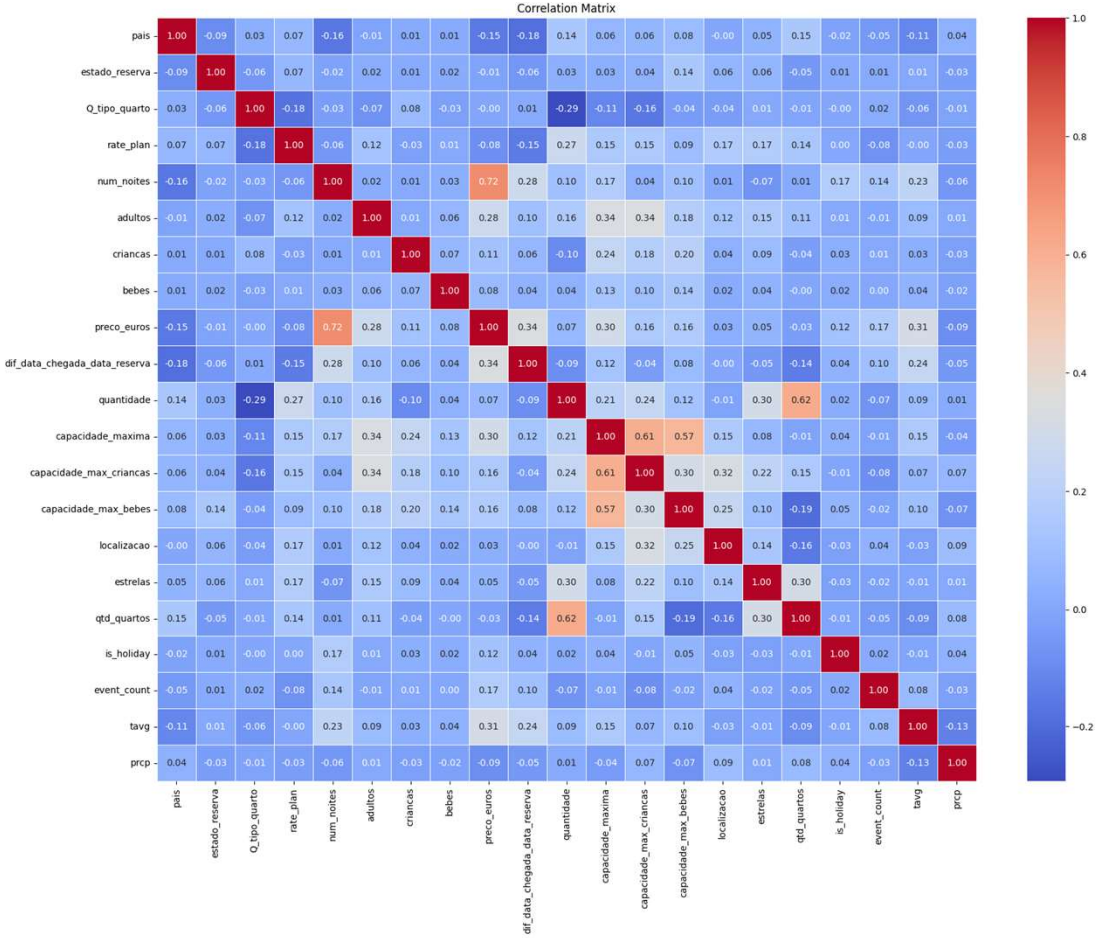
Integração dos dados

- Dados da e-GDS foram exportados em csv e lidos em spark.
- Junção dos dados com base nas colunas em comum
  - O hotel 309 não têm nenhuma reserva
- Juntando os feriados criamos a coluna “is_holiday” que define se há um feriado durante a reserva.
- Juntando os eventos criamos a coluna “event_count”.
- Juntando a meteorologia calculamos a média de temperatura e a temperatura máxima e mínima durante os dias da reserva.
- Além da temperatura juntamos a coluna prcp (precipitação)

Análise exploratória depois da integração

Análise exploratória





Map Reduce – Reservas por hotel

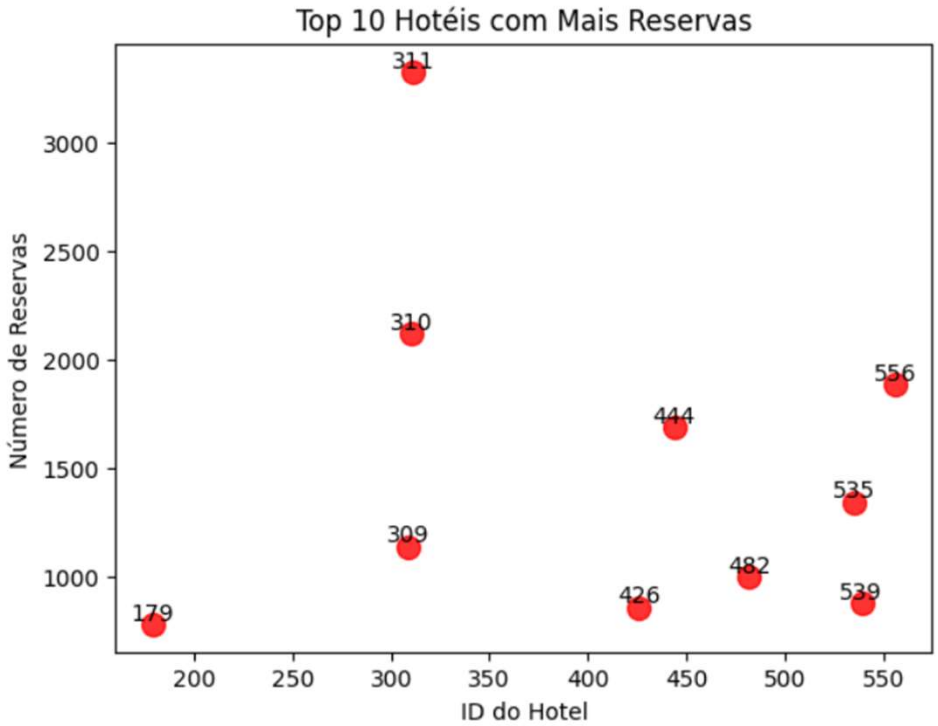


Figura 8.3: Top 10 Hotéis com mais reservas

Map Reduce – Média de noites por origem

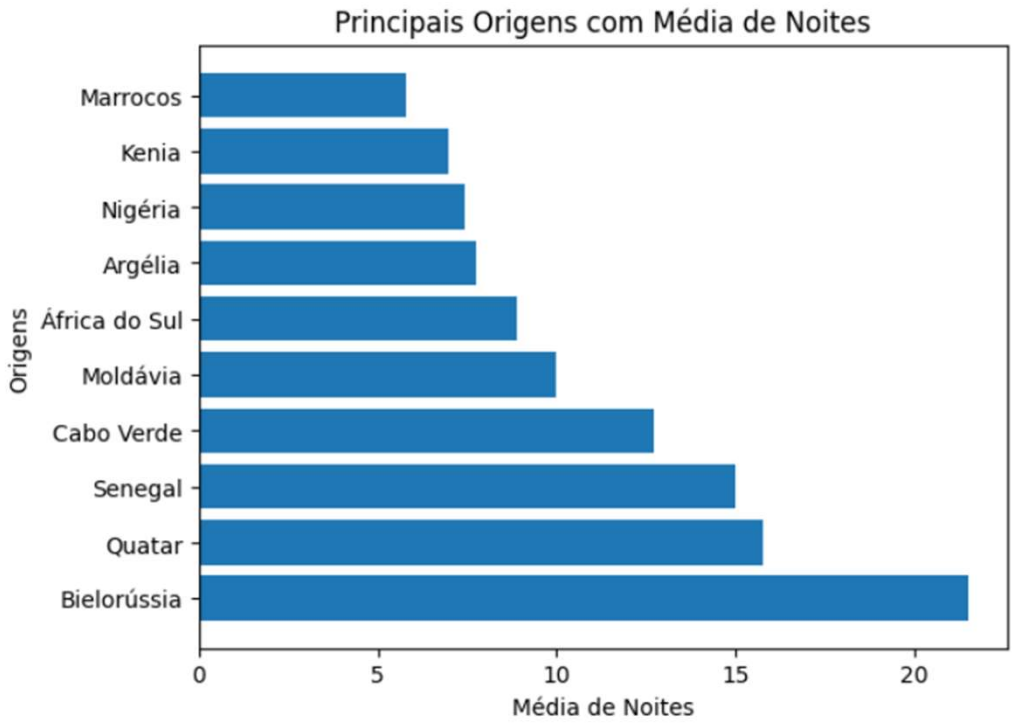


Figura 8.4: Média de noites por origem

Map Reduce – Preço médio por noite e por hotel

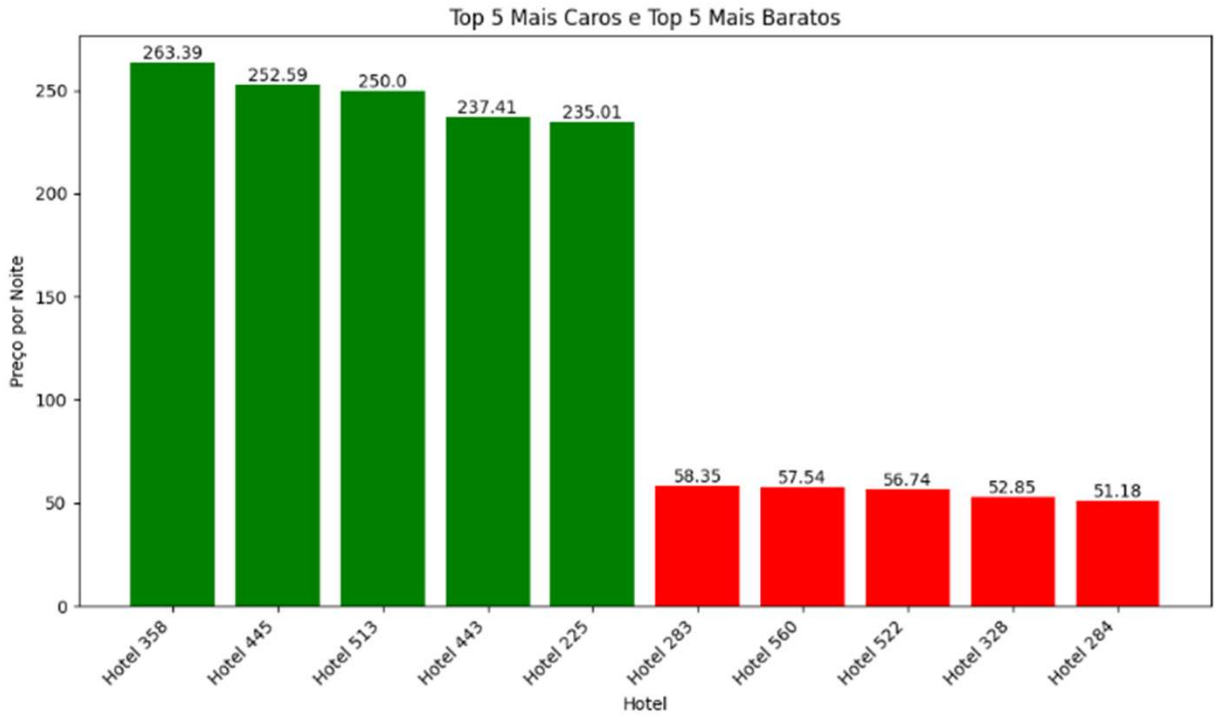
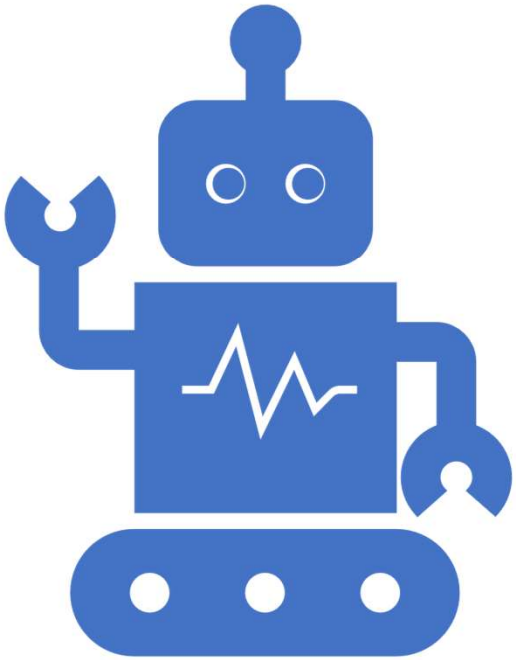


Figura 8.5: Top 5 hoteis mais caros e Top 5 mais baratos



Machine Learning

Machine Learning

- Criação de novas colunas com a temporada das reservas
- Criação da coluna cancelamento
- Criação da coluna estação de chegada

Função de Classificação de Atributos
Feature Ranking

Coluna	Importância
room ID	0,2386
qtd_quartos	0,1976
hotel_ID	0,1535
temerature_avg	0,1117
adultos	0,0769
estrelas	0,0706
num_noites	0,0563
capacidade_maxima	0,0385
criancas	0,0267
event_count	0,0118
temporada	0,0116
is_holiday	0,0061

Machine Learning - Preço

Tabela 9.1: Tabela de Resultados GBTRegressor

room_ID	preco_noite_adulto	prediction
81	28.5	31.29484738379246
81	29.5	29.08658871842678
81	31.25	28.345357587874094
81	28.5	25.498833181179428
81	27.0	26.84495168929202
85	62.5	67.9726062040213
85	67.0	74.11553050981094
190	57.0	67.6776023950932
190	45.5	53.25816632902684
85	72.5	67.9726062040213
85	107.0	84.1241782911615
190	100.0	122.15293974790283

RMSE: 17.54049

Tabela 9.2: Tabela de Resultados RandomForestRegressor

room_ID	preco_noite_adulto	prediction
81	28.5	58.546431042678556
81	29.5	52.58288062454881
81	31.25	58.546431042678556
81	28.5	43.03895199281975
81	44.0	82.93799806188757
85	62.5	62.55340415968901
85	67.0	58.693005417616746
190	57.0	60.23613863543294
190	45.5	51.436251808719476
85	72.5	60.636429933126806
85	107.0	93.97093037857026
190	100.0	97.01677646841392

RMSE: 21.683

Machine Learning - Cancelamentos

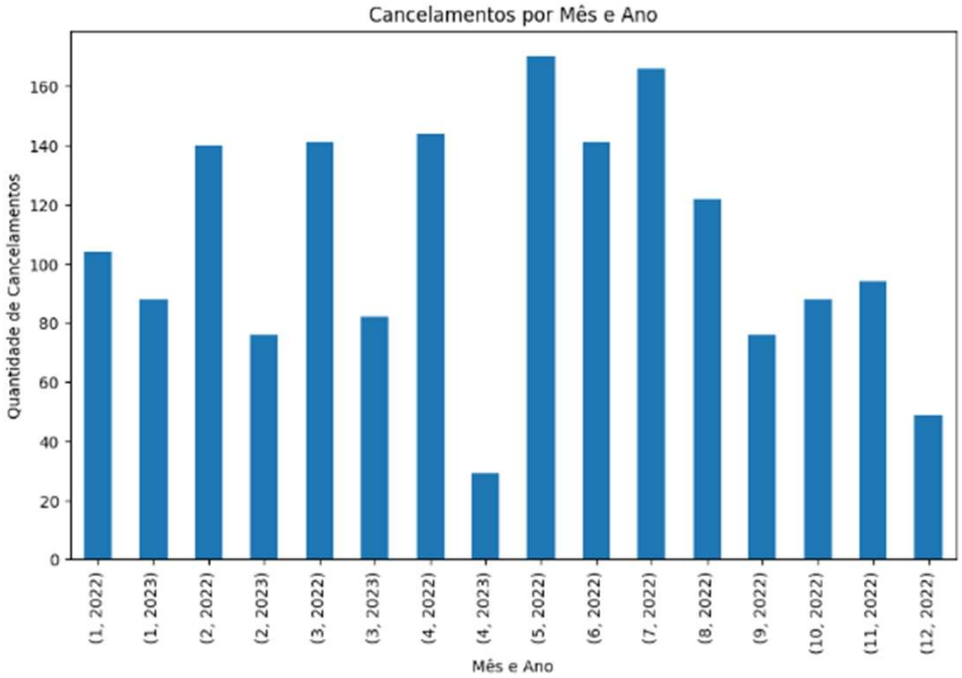


Figura 9.1: Cancelamentos Geral de Reservas

Machine Learning - Cancelamentos

- Colunas iniciais:
  - hotel_ID
  - adultos
  - capacidade_maxima
  - qtd_quartos
  - Is_holiday
  - pais
  - preco_noite
  - temporada
  - tipo_quarto
  - localizacao

Machine Learning - Cancelamentos

Colunas	Accuracy
Colunas iniciais	0,51
Adição “estrelas”	0,5886
Adição “dif data chegada data reserva”	0,6153
Adição “rate plan” (removida posteriormente)	0,606
Adição “num noites” (removida posteriormente)	0,6115
Adição “is holiday” (removida posteriormente)	0,6099
Adição “crianças” (removida posteriormente)	0,6146
…	…

Machine Learning - Cancelamentos

Tabela 9.3: Tabela de Resultados

Hotel	**Mês**	***accuracy***	**Cancelamentos**	**Reservas**
444	1	0.5348	3	115
444	7	0.6051	8	132
309	7	0.6690	7	60

RandomForestClassifier

Conclusões

Conclusões

- Dificuldade na procura e integração de dados externos que se relacionem com as reservas
- Mudança frequente do conjunto de dados ao longo do projeto
- Visualizações gráficas para ajudar na compreensão dos dados
- Análise extensiva dos conjuntos de dados
  - Tratamentos de grande quantidade e diversidade de dados
  - Dados duplicados e novas colunas

Conclusões

- Resultados de ML foram limitados devido à natureza restrita do conjunto de dados (apenas 1 ano completo)
- Conjuntos de dados mais completos e confiáveis pode levar a resultados mais precisos e significativos
- Exploração mais profunda de técnicas de ML e incorporação de outras variáveis relevantes
- Análise mais aprofundada e uma maior diversidade de dados poderiam levar a resultados mais robustos

TEAD - TP

João Bragança – 8190555

José Fernandes – 8190239

Pedro Afonso – 8090457