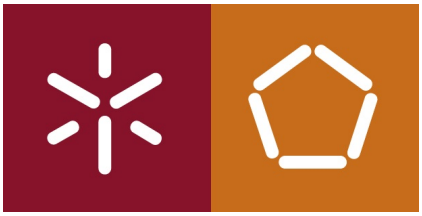

Universidade do Minho
Mestrado Integrado em Engenharia Informática
4ºano - 1º Semestre

Análise de Dados

Trabalho Prático

Grupo 03

A83732 – Gonçalo Rodrigues Pinto
A80791 – João Diogo Mendes Teixeira Da Mota
PG42839 - José Gonçalo Macedo Costa
A84829 - José Nuno Martins da Costa

24 de Janeiro de 2021

Conteúdo

Lista de Figuras

1 Introdução

A manipulação e exploração de dados alcançou um patamar de excelência com as plataformas de processamento analítico. Existem, hoje, na área de *Business Analytics*, inúmeras plataformas que disponibilizam a qualquer área a possibilidade de, através de dados históricos dessa organização, a análise de determinados aspectos passíveis de ser explorados e melhorados em prol da organização em causa. Por este motivo, a existência da presente Unidade Curricular de Análise de Dados, no perfil de Engenharia do Conhecimento do curso de Engenharia Informática da Universidade do Minho, é vista como essencial, pelo seu programa lecionado ter integrada a área de *Business Intelligence*, dada a sua importância no apoio de decisões e justificação de investimentos em qualquer área hoje em dia.

O presente trabalho enquadra-se nessa Unidade Curricular e a sua elaboração pretende atribuir aos alunos as competências estudadas ao longo da mesma. Estas competências dividem-se na utilização de ferramentas de *Business Intelligence*, e em toda a fase necessária para preparação de um sistema de suporte a esta área. Para a exploração destes dois temas centrais, foi dada a liberdade de escolha do objecto de estudo, tendo o grupo selecionado uma fonte de dados que poderia perfeitamente se enquadrar no tema requerido.

Espera-se, durante e após a realização do projecto proposto, a consolidação de diversos tópicos abordados e discutidos durante as aulas da Unidade Curricular, assim como a capacidade de aplicação de diversas estratégias e métodos em situações reais.

2 Descrição do problema

De modo a desenvolver um sistema de *Data Warehousing*, bem como um sistema de *Business Intelligence* para suporte à decisão, considerou-se o seguinte:

- Realizou-se um trabalho de análise, planeamento, e implementação, tendo como base um *dataset* público à escolha do grupo;
- A escolha do *dataset* público compreendeu a escolha de uma área de negócio, permitindo o desenvolvimento de indicadores relevantes para o caso de estudo;
- Foi desenvolvido um sistema de povoamento inicial e as estruturas necessárias à sua atualização de forma incremental e/ou diferencial;
- Foi proposta a utilização de uma das seguintes plataformas de desenvolvimento de **Business Intelligence**: *Microsoft Power BI Desktop* ou *Tableau Desktop*, optando-se pela plataforma *Tableau*.

3 Análise do Dataset

Numa primeira fase, foi realizada uma pesquisa sobre a utilidade do *dataset* para compreensão e construção de um sistema de base de dados multidimensional. Assim, salienta-se o patamar de excelência alcançado pela manipulação e exploração expedita de dados com as plataformas de processamento analítico.

Os agentes de decisão empresariais têm atualmente, à sua disposição, inúmeras ferramentas que permitem manipular os dados que dispõem segundo as suas mais diversas perspetivas de negócio, posicionando-se e deslocando-se em espaços multidimensionais de dados, na procura de novos elementos que possam suportar as suas decisões. Requer-se, em alguns casos, um passo mais direcionado à automatização e otimização dos processos de negócio e, consequentemente, das suas ações de tomada de decisão. Para este fim, pode recorrer-se a sistemas de *Business Analytics*.

Desta forma, pretendeu-se estudar e aplicar técnicas de *Business Analytics*, num contexto de aplicação empresarial, na tentativa de encontrar padrões de dados e respetivos relacionamentos no *dataset*, clarificando a ocorrência de determinados resultados. Para além disso, foram realizadas previsões de resultados futuros para o contexto específico do negócio, com a descoberta de conhecimento para a implementação de novos processos de negócio e de tomadas de decisão.

3.1 Dataset selecionado

De forma a escolher a área de análise, o grupo selecionou alguns temas que considerou interessantes. Um dos temas que, desde logo, saltou à vista foi o tema de viagens. Como tal, e uma vez que existem diversas plataformas para escolher um destino de viagem, decidiu-se desenvolver o projeto em torno desta área.

Uma dessas plataformas mais populares e em corrente crescimento é a **Airbnb**, que apresenta instalações desde o mais simples quarto até a uma luxuosa vivenda. Desde a sua criação, em 2008, o *Airbnb* disparou em popularidade como alternativa aos hotéis. O serviço de hospitalidade *online* é uma área de negócio em constante expansão, que permite a exploração de lugares autênticos, sendo apenas necessário entrar em contacto com o anfitrião dos respetivos lugares.

Dados disponibilizados pela **Airbnb** revelam a existência de 2.9 milhões de *hosts* em todo o mundo, em 2020. Desses, 14.000 são criados por mês. Atualmente, existem mais de 7 milhões de anúncios na plataforma, com informações de 100.000 cidades pertencentes a uma totalidade de 220 países.

O sucesso da *Airbnb* requer a constante compreensão e aperfeiçoamento do seu negócio, de forma a preservar a sua avaliação de 100 mil milhões de dólares. Para este efeito, é indispensável a análise contínua dos seus dados. Neste projeto, foram analisados estes dados, de forma a perceber o sucesso desta plataforma. A pesquisa sobre estes dados levou ao encontro da página *Inside Airbnb* [2]. Esta página apresenta um conjunto independente, e não comercial, de ferramentas e dados que permitem a exploração do modo como o *Airbnb* é utilizado em escala mundial.

A perceção do seu impacto exigiu a reunião de grandes quantidades de dados, que se encontram divididos por cidades. Assim sendo, recorreu-se à plataforma *OpenDataSofts* [1], que fornece vários *datasets* públicos. Esta plataforma possui um *dataset* livre que agrupa várias cidades com diversas informações acerca de uma determinada hospedagem. Este *dataset* fornece métricas indispensáveis para a compreensão do modo como o *Airbnb* é utilizado, permitindo uma análise objetiva sobre estes dados.

3.2 Informação presente no *dataset*

O *dataset* selecionado apresenta diversas entidades e respetivas descrições. A análise abrangente deste conjunto de dados exigiu a organização destas entidades e campos.

- **Estabelecimento** - Condições do estabelecimento onde os clientes podem ficar hospedados. Possuem atributos como um identificador, hiperligações para várias fotografias, um nome, um sumário, descrição, espaço, experiência oferecida, regras, tipo de propriedade, tipo de quarto, quantos hóspedes alberga e quantas casas de banho e quartos contêm. É, ainda, mencionado o número de camas e respetivo tipo. Além disto, são descritas as principais comodidades que possuem.

- **Anfitrião** - Responsável pelo estabelecimento. Caracterizado por um identificador, um nome, ligações para o seu perfil, a data desde que é anfitrião, a sua localização, o tempo que demora a responder, bem como uma classificação e ainda a taxa de aceitação. Além do mais, possui meios de contacto e uma breve descrição pessoal. É, também, possível verificar o número de estabelecimentos que possui, tipo de verificação e cancelamento exigido. Pode ainda conter informações adicionais.

- **Localização** - Local do estabelecimento. Descrito por um bairro, cidade, estado, código de postal, o mercado onde opera e país, bem como a sua latitude e longitude. Apresenta uma visão geral do bairro, notas adicionais, trânsito e formas de acesso.

- **Pontuação** - Várias pontuações médias que são dadas ao estabelecimento. Possui o total de *reviews* realizadas e as datas da primeira e última *review*. De forma mais detalhada, apresenta pontuações sobre os vários parâmetros relativos à estada num destes estabelecimentos (limpeza, localização, entre outros).

- **Hospedagem** - Exibe a afluência do estabelecimento, o número de clientes que se encontram atualmente hospedados no estabelecimento e o número de hóspedes que pode, ainda, albergar. Para além disto, são encontrados detalhes sobre a disponibilidade imediata até 365 dias (mensal, trimestral, semestral) e os diversos preços, apresentados por noite, semana e mês; o valor do depósito de segurança e o preço de limpeza.

3.3 Dados

Dada a extensão do *dataset* selecionado, foram definidos apenas os atributos que não apresentam frequentemente valores nulos, facilitando a análise do seu povoamento.

- **Estabelecimento**: Dados consistentes, embora pouco legíveis. Algumas colunas têm valores nulos predominantes e demasiado extensos;
- **Anfitrião**: Dados consistentes. Evidenciam-se colunas com valores nulos predominantes, bem como colunas com demasiado detalhe;
- **Localização**: Dados legíveis. Algumas colunas têm valores nulos predominantes. Outras colunas apresentam repetições.
- **Pontuações**: Dados organizados e legíveis. São apresentadas colunas com valores nulos predominantes. Várias colunas encontram-se em minoria.
- **Hospedagem**: Dados organizados e legíveis, apesar de existirem diversas colunas com valores nulos predominantes.

3.4 Vertente de análise selecionada

Apresentam-se algumas possíveis vertentes de análise para a listagem dos **Airbnb**:

- **Relatórios sobre as hospedagens**: Este relatório pode ser usado para detetar variâncias nos preços e condições de estadia, mediante o estabelecimento que possuem. Terá em consideração as diversas pontuações que atribuídas, assim como o tipo de cancelamento e experiência que proporciona;

- **Relatório sobre as pontuações**: Relatório para deteção dos tipos de pontuações dadas conforme a localização, tipo de estabelecimento ou anfitrião. Útil para análise de fatores a serem melhorados nos diferentes tipos de estabelecimento, e averiguar as condições mais valorizadas pelos hóspedes;

- **Relatório ao nível do anfitrião**: Este relatório permite a análise dos estabelecimentos dos anfitriões, possibilitando o seu aperfeiçoamento, quer oferecendo melhores condições, quer a repensar preços e/ou localizações;

- **Distribuição das localizações dos estabelecimentos**: Relatório conveniente para encontrar estabelecimentos a nível global, localizar hospedagens onde existe maior oferta, entre outros.

A gestão eficiente do serviço de hospitalidade *online* é conseguida com o estudo das potenciais áreas de negócio. Optou-se pelo estudo de uma vertente focada nos relatórios das métricas e análise de algumas relações com as diversas pontuações. Assim, estudou-se:

- **Relatórios sobre as hospedagens**: Pode ser estabelecida uma relação entre o preço praticado, o número de membros e a experiência oferecida, com as condições e localização, tal como o a influência do anfitrião;

- **Relatórios simples ao nível do anfitrião, localização, tipo de propriedade**: Permite conhecer os anfitriões de renome, localizações mais comuns de estabelecimentos e, ainda, os tipos mais predominantes nestes serviços;

- **Relatórios simples de pontuações:** Permite o analisar as melhores *reviews*, bem como as condições preferenciais.

4 Data Warehouse

4.1 Funcionamento do sistema

O processo de ETL (*Extract-Tranform-Load*) é fundamental para qualquer iniciativa de *Data Warehousing*. A figura 1 descreve qualquer processo de ETL e respetiva aplicabilidade. No presente trabalho prático, foram extraídos dados provenientes de uma única fonte de informação.

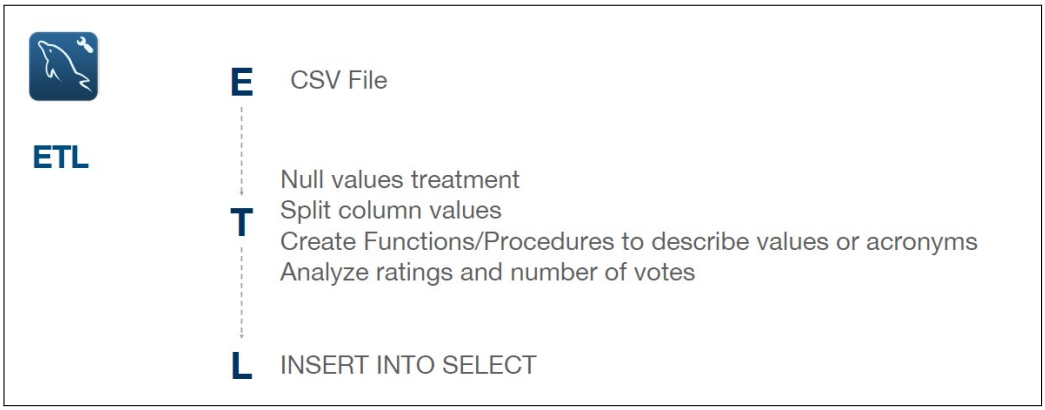

Figura 1: Processo de ETL.

Assim, o ETL foi visto como uma ligação entre os dados e o *Data Warehouse*. Este processo teve início com a extração dos dados da fonte de informação para uma área de retenção. Nesta primeira etapa, não foi extraída a totalidade dos dados das bases de dados, mas apenas os dados pertinentes e para posterior análise. Posteriormente, realizou-se a limpeza, consolidação e conformidade dos dados. Neste passo, foram os dados foram transformados e adaptados para o *Data Warehouse* final. Por fim, inseriu-se os dados preparados no *Data Warehouse*, garantindo a qualidade e valor dos dados para posterior para análise e apoio de decisões.

4.2 Seleção de Dados

De forma a recolher informação útil e relevante, foi necessário definir quais os dados a ser analisados, bem como as informações que a extrair. Dada a extensão do *dataset*, foram apenas consideradas hospedagens cuja última pontuação foi atribuída em 2017.

Analisando os dados do **Airbnb**, verifica-se a existência de alguns elementos fundamentais para a compreensão e análise das hospedagens:

- Preço por noite;
- Preço por semana;
- Preço por mês;
- Depósito de segurança;
- Taxa de limpeza;
- Número de hóspedes;
- Número extra de hóspedes;
- Mínimo de noites;
- Disponibilidade (anual);
- Politica de Cancelamento;
- Características;
- Tipo de Experiência oferecida;

Posto isto, foi elaborado um conjunto de questões a serem respondidas.

1. Preço
    1.1. Como varia temporalmente o preço comparado com o número médio de hospedagens dos anfitriões?
    1.2. Como varia o preço por tipo de quarto?
    1.3. Quais as cidades que praticam preços mais elevados?
2. Preço por semana e por mês
    2.1. O preço semanal está diretamente relacionado com o preço mensal praticado em cada tipo de propriedade?

3. Depósito de segurança

    3.1. Como varia o preço do depósito de segurança nas principais cidades?

4. Taxa de limpeza

    4.1. O tipo de quarto, juntamente com o tipo de cama, reflete um aumento da taxa de limpeza?

5. Número de hóspedes e Número extra de hóspedes

    5.1. Os 20 serviços mais comuns são adequados ao número de hóspedes e pessoas extra que pode albergar?

6. Mínimo de noites

    6.1. Existe alguma relação entre o número mínimo de noites e o tipo de quarto a hospedar, nas três principais propriedades?

7. Disponibilidade (anual)

    7.1. Como varia trimestralmente a disponibilidade consoante o número médio de hospedagens dos anfitriões?

8. Politica de Cancelamento

    8.1. Qual o tipo de cancelamento mais praticado?

    8.2. Qual o tipo de política de cancelamento que cada país pratica?

    8.3. Das cidades onde existem pelo menos 2000 anfitriões registados, como são as políticas de cada uma delas?

    8.4. Como é a distribuição da política de cancelamento para cada tipo de quarto?

9. Tipo de Experiência oferecida

    9.1. Qual a experiência mais predominante?

    9.2. Quais as cidades com pelo menos cinco registos que proporcionam experiências?

    9.3. Existe alguma tendência em fornecer experiências ao longo dos anos?

    9.4. O tipo de propriedade influencia a experiência?

10. Características

    10.1. Quais as características mais comuns?

11. Outras

    11.1. Qual o top de dez de serviços por tipo de quarto?

    11.2. Qual a distribuição dos dez serviços mais comuns pelas dez cidades com mais hospedagens?

    11.3. Quais os países com mais hospedagens?

    11.4. Como é a afluência de registos de novas hospedagens?

    11.5. Qual é o tempo de resposta mais predominante do top 20 da origem de cada anfitrião?

    11.6. O tempo de resposta influencia a pontuação ao nível da resposta?

    11.7. Qual é a média da quantidade de estabelecimentos por país?

    11.8. Qual é o tipo de verificação mais utilizado?

    11.9. Como é a distribuição do top cinco de verificações por tipo de quarto?

    11.10. Qual o tipo mais predominante de quarto por tipo de propriedade?

    11.11. Qual o tipo mais predominante de cama por tipo de quarto?

    11.12. Existirá uma cama por quarto?

    11.13. Quais são as propriedades com maior pontuação e de que forma reflete um grande número das mesmas em cada mês?

    11.14. Qual a experiência a nível geral por trimestre?

    11.15. A exatidão que a página de anúncio representa o espaço está relacionado com a pontuação da relação qualidade/preço?

    11.16. A experiência de um anfitrião e o seu tempo de resposta influencia pontuações referentes?

    11.17. Como se sentiram os hóspedes em relação à vizinhança em cada país, sabendo que é uma localização exata?

    11.18. O tipo de quarto influencia a pontuação referente à limpeza?

4.3 Modelo Dimensional

Foi, então, criado um modelo dimensional com base nos dados apresentados.

4.3.1 Dimensões e Factos

A definição das dimensões e factos constituintes do modelo dimensional teve por base a interpretação das questões previamente levantadas. Deste modo, e sabendo que as dimensões são utilizadas com o objetivo de filtrar e categorizar os factos (medidas), foram consideradas as seguintes dimensões:

1. **Data**: Detalhes sobre a data (dia, mês e ano);
2. **Cama**: Informações sobre o tipo de cama;
3. **Política de Cancelamento**: Informações sobre o tipo de cancelamento;
4. **Cidade**: Informações sobre a cidade;
5. **País**: Informações sobre o país;
6. **Experiência Oferecida**: Informações sobre o tipo de experiência;
7. **Localização do anfitrião**: Informações sobre a localização do anfitrião;
8. **Bairro**: Informações relativas ao bairro;
9. **Tipo de Propriedade**: Informações sobre o tipo de propriedade;
10. **Tempo de resposta do *host***: Informações sobre o tipo de resposta do anfitrião;
11. **Tipo de quarto**: Informações sobre o quarto;
12. **Serviço**: Informações sobre os serviços fornecidos;
13. **Característica**: Informações sobre as características;
14. **Verificação**: Informações sobre o tipo de verificação que o anfitrião possui;
15. **Proprietário**: Informações sobre o identificador, data de registo, origem, tempo de resposta, pontuação referente à resposta e número de hospedagens listadas que possui;

16. **Estabelecimento**: Informações relativas ao tipo de propriedade, quarto, número de pessoas que alberga, número de casas de banho, quartos e camas, bem como o tipo de cama;

17. **Localização**: Informações sobre o bairro, cidade e país, juntamente com latitude e longitude do anfitrião;

18. **Pontuações**: Informações a respeito das pontuações atribuídas pelos hóspedes ao nível de quantidade, data das primeira e última pontuações, pontuações de experiência no geral, limpeza, exatidão, *check-in*, comunicação, localização, relação qualidade/preço e ainda o número de *reviews* mensais;

19. **Serviços**: Associação entre uma instalação e um serviço;

20. **Verificações**: Associação entre uma verificação e um anfitrião.

De igual forma, foram identificados os factos que se seguem.

1. **Hospedagem**: Contém atributos não descritivos sobre cada hospedagem, tais como o preço (diário, semanal e mensal), depósito de segurança, taxa de limpeza, número de hóspedes, número extra de hóspedes, mínimo de noites, disponibilidade (por ano), politica de cancelamento, características e tipo de experiência oferecida.

4.3.2 Esquema

Tendo por base as dimensões e factos identificados, surgiu um modelo dimensional em **floco de neve**, dada a existência de uma única tabela de factos. Para além disso, sentiu-se a necessidade de representar várias hierarquias de dimensões, permitindo uma melhor separação das entidades a tratar.

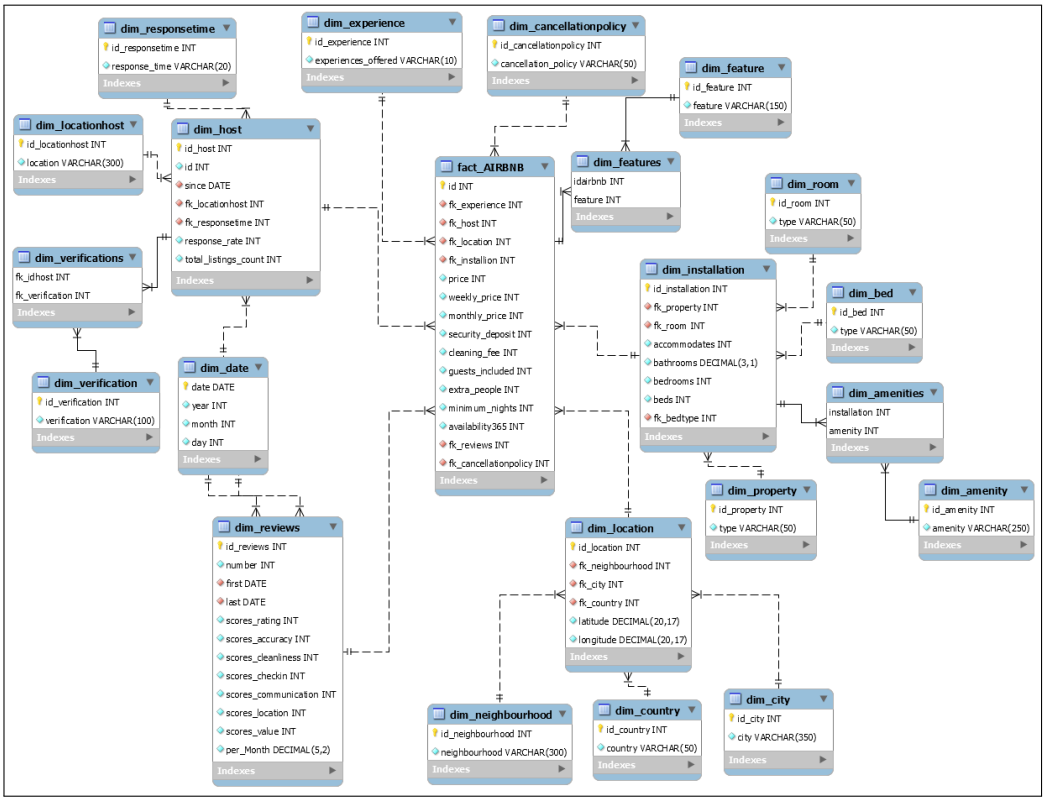

Figura 2: Modelo Dimensional.

4.4 *ETL*

Uma vez definido e implementado o *Data Warehouse*, foi necessário realizar o povoamento inicial com os dados presentes no *dataset*. Neste, filtrou-se as colunas indispensáveis à resposta das questões colocadas. Além disto, foram considerados apenas os dados cuja data da última pontuação tenha sido em 2017. Contudo, o *dataset* permaneceu muito extenso, originando problemas durante o seu processamento. Numa primeira fase, realizou-se o carregamento dos dados com as ferramentas que o sistema de gestão de bases de dados *MySQL Workbench* fornece. No entanto, devido à extensão do *dataset*, o carregamento dos dados a bruto para este sistema chegou a demorar quase um dia, atrasando muito este processo. De forma a solucionar este problema, optou-se por desenvolver um *parser*, na linguagem de programação *JAVA*. Neste *parser*, implementou-se a extração, tratamento e carregamento dos dados para um ficheiro de povoamento SQL, posteriormente executado no SGDB, permitindo o povoamento do *Data Warehouse* construído.

Para suportar as fases do *ETL*, foi criada uma "área de retenção"em memória, com auxílio de estruturas e classes que a linguagem utilizada permite. Trata-se de uma área de transição dos dados entre o *dataset* e o *Data Warehouse* final. Assim, o processo de preenchimento resume-se a:

1. Extrair dados do *dataset* para a Área de Retenção;
2. Transformar os dados na Área de Retenção;
3. Conciliar e extrair os dados da Área de Retenção para o *Data Warehouse* final.

4.4.1 Extração e Transformação

Esta fase correspondeu ao estágio de transformação dos dados presentes no processo de ETL como *Transform*, correspondente à limpeza, operações de agregação e filtragem.

Este processo teve início com a armazenação do *dataset* num *buffer*. Posteriormente, iterou-se sobre este *buffer*, lendo linha a linha e tratando da separação destas pelo carater ";". Desta forma, obteve-se uma lista de palavras, onde cada posição corresponde aos parâmetros previamente apresentados. Por questões de simplificação, este processo foi manipulado com campos textuais, sem realização de qualquer conversão.

```
BufferedReader in = new BufferedReader(
    new FileReader("csv\\Information_metrics_for_listings.csv") );
String line;
String[] campos;
while ((line = in.readLine()) != null) {
    campos = line.split(";");
            ...
}
```

Em seguida, realizaram-se verificações com o intuito de **tratar os valores nulos** presentes nos diferentes parâmetros, de modo a alterar a sua representação no *Data Warehouse*. Procedeu-se à aplicação dos seguintes critérios:

- Valores nulos em atributos textuais são marcados com a palavra *"unknown"*;
- Valores nulos presentes em atributos numéricos são atribuídos como o valor -1;
- Valores nulos em atributos que representam datas são atribuídos com uma data que não se encontra associada a nenhuma hospedagem (optou-se pelo dia *"01/01/1998"*);

```
...
String hostID = campos[2];
if (hostID.equals("")) { hostID = "-1"; }
String hostSince = campos[3];
if (hostSince.equals("")) { hostSince = "01/01/1998"; }
String hostLocation = campos[4];
if (hostLocation.equals("")) { hostLocation = "unknown"; }
...
```

4.4.2 Carregamento

Numa segunda fase, realizou-se o carregamento sequencial dos dados. Isto é, averiguou-se se os elementos mais simples (por exemplo, o nome de uma cidade ou o tipo de cama) já tinham sido carregados previamente para memória na estrutura criada para armazenar valores deste tipo. Para estas estruturas, foram utilizadas predominantemente coleções onde a sua chave é constituída pelos elementos armazenados e, no respetivo valor, o número desse elemento na estrutura em causa. Segue-se um exemplo do processamento do elemento que traduz o nome de uma cidade.

```
int cities_counter = 1 ;
HashMap<String, Integer> cities = new HashMap<>();
...
if (!cities.containsKey(city)) {
    cities.put(city, cities_counter++);
}
...
```

Os atributos passíveis de serem compostos por mais do que um elemento (como é o caso dos Serviços, Verificações e Características) foram tratados de forma independente. Os elementos de cada atributo foram divididos pelo carater ",". Desta forma, foi criada uma nova lista de elementos de um dado tipo e, posteriormente, realizou-se o mesmo processo anteriormente apresentado para um elemento simples.

No excerto que se segue, está representado o processo de separação das verificações.

```
int verification_counter = 1 ;
HashMap<String, Integer> verifications = new HashMap<>();
...
for (String s : verification_fields) {
    int id_verification = verifications.get(s);
    int id_host = hosts.get(key_host).getId();
    verificationsArrayList.add(
                new Verifications(id_host, id_verification));
}
```

Subsequentemente, foram criadas diversas classes que representam dimensões compostas. Ou seja, dimensões formadas por referência a elementos simples previamente carregados. Assim, é possível armazenar a informação nesta "área de retenção"com a mesma lógica previamente apresentada. No

entanto, ao invés de o valor ser o contador, é um objeto criado. Para tal, construiu-se de uma chave composta pelos elementos da mesma. Apresenta-se, em seguida, o exemplo para o caso da representação de um determinado anfitrião.

```
class Host {
    private int id;
    private String id_host;
    private String since;
    private int fk_locationhost;
    private int fk_responsetime;
    private String response_rate;
    private String total_listings_count;
}

int hosts_counter;
HashMap<String, Host> hosts = new HashMap<>();
...
String key_host = "" + hostID + ";" + hostSince + ";"
    + hostLocation + ";" + hostResponseTime + ";"
    + hostResponseRate + ";" + hostTotalListingsCount + ";" + id;
if (!hosts.containsKey(key_host)) {
    Host l = new Host(hosts_counter++,
        hostID,
        hostSince,
        host_location.get(hostLocation),
        host_responsetime.get(hostResponseTime),
        hostResponseRate, hostTotalListingsCount);
    hosts.put(key_host, l);
}
```

Em seguida, foram carregados os elementos compostos e simples. Assim sendo, a fase que se segue focou-se no estabelecimento das devidas correspondências entre estes, como é o caso dos Serviços com Estabelecimento e Verificações com Anfitriões. Este processo baseou-se na iteração pela nova lista de elementos simples e associação à chave criada. Evidencia-se, em seguida, a associação das verificações:

```
ArrayList<Verifications> verificationsArrayList =
                                new ArrayList<>();
...
for (String s : verification_fields) {
```

```
    int id_verification = verifications.get(s);
    int id_host = hosts.get(key_host).getId();
    verificationsArrayList.add(
        new Verifications(id_host, id_verification));
}
```

Por fim, foi necessário representar a informação dos factos numa nova classe com os respetivos elementos. Posto isto, restou agrupá-la com a informação previamente processada.

```
public class Facts {
    private int id;
    private int experience;
    private int host;
    private int location;
    private int installation;
    private String price;
    private String weekly_price;
    private String monthly_price;
    private String security_deposit;
    private String cleaning_fee;
    private String guests;
    private String extras;
    private String minimum_nights;
    private String availability;
    private int reviews;
    private int cancellationpolicy;
}
...
ArrayList<Facts> factsArrayList = new ArrayList<>();
...
factsArrayList.add(new Facts(
    Integer.parseInt(id)
    , experiences.get(experiencesOffered)
    , hosts.get(key_host).getId()
    , locations.get(key_location).getId()
    , installations.get(key_installation).getId()
    , price, weeklyPrice, monthlyPrice, securityDeposit
    , cleaningFee, guestsIncluded, extraPeople, minimumNights
    , availability, reviews.get(key_review).getId()
    , cancellations_policy.get(cancellationPolicy)
));
```

Este processo foi, ainda, realizado ao nível dos Serviços e Verificações, interligando o facto com as suas características.

Após processamento do ficheiro, iterou-se sobre cada estrutura implementada, obtendo os seus valores armazenados para escrita do ficheiro de povoamento.

```
PrintWriter out = new PrintWriter(
    new BufferedOutputStream(
        new FileOutputStream("sql\\settlement.sql",true)));
...
...
for (Facts a : factsArrayList) {
    String insert = "insert into fact_AIRBNB values ("
        + a.getId()
        + "," + a.getExperience() + "," + a.getHost()
        + "," + a.getLocation() + "," + a.getInstallation()
        + "," + a.getPrice() + "," + a.getWeekly_price()
        + "," + a.getMonthly_price()
        + "," + a.getSecurity_deposit()
        + "," + a.getCleaning_fee() + "," + a.getGuests()
        + "," + a.getExtras() + "," + a.getMinimum_nights()
        + "," + a.getAvailability() + "," + a.getReviews()
        + "," + a.getCancellationpolicy()
        + ");";
    out.println(insert);
    out.flush();
}
```

O ficheiro de povoamento foi executado como um *script*, de forma a povoar o *Data Warehouse*. É de realçar que, para o povoamento da dimensão *Data*, foi utilizado um procedimento que insere todas as datas compreendidas num intervalo de tempo, aquando da sua invocação (realizada no início do *script*).

4.5 Testes

A validação do correto carregamento dos dados foi concretizada com a implementação de diversos procedimentos. Estes procedimentos recebem o identificador de uma determinada hospedagem e devolve a informação - total ou parcial - respetiva (Anfitrião, Estabelecimento, Localização, Pontuações) ou, ainda, as diferentes componentes que a compõem (Serviços, Verificações, Características).

```
DELIMITER $$
CREATE PROCEDURE GetHost(IN id INT)
BEGIN
SELECT f.id, h.id, h.since, lh.location, r.response_time,
h.response_rate, h.total_listings_count
FROM fact_airbnb f, dim_host h, dim_locationhost lh,
dim_responsetime r
WHERE f.fk_host = h.id_host
AND h.fk_locationhost = lh.id_locationhost
AND h.fk_responsetime = r.id_responsetime
AND f.id = id ;
END $$

DELIMITER $$
CREATE PROCEDURE GetVerifications(IN id INT)
BEGIN
SELECT f.id, v.verification
FROM fact_airbnb f, dim_host h, dim_verifications vv,
dim_verification v
WHERE f.fk_host = h.id_host AND h.id_host = vv.fk_idhost
AND vv.fk_verification = v.id_verification AND f.id = id ;
END $$
```

A criação adicional de uma *view* permitiu a visualização da informação presente no *dataset*, com exceção das informações de Serviços, Verificações e Características.

4.6 Atualização incremental e diferencial

Um *Data Warehouse* deve ser atualizado regularmente, de modo a garantir que as informações derivadas deste estejam atuais. Assim, implementou-se um sistema de atualização diferencial e incremental. Desta forma, serão apenas carregados os dados que não estão presentes no *Data Warehouse*, ou seja, corresponde à inserção de novas informações ou a substituição das mesmas. Como não é considerado o histórico de dados, qualquer dado modificado na Base de Dados original é também modificado no *Data Warehouse*. Por exemplo, se uma hospedagem mudar a data da última pontuação, o *Data Warehouse* não irá manter um histórico de pontuações, e apenas irá considerar os dados atuais.

Assim sendo, dividiu-se este processo, desenvolvendo para o efeito duas aplicações, uma responsável pela inserção de nova informação e outra encarregue de atualizar o *Data Warehouse*. Em cada uma destas aplicações criou-se um *dataset* com a nova informação. Ambas aplicações foram desenvolvidas na linguagem de programação JAVA, uma vez que o povoamento do *Data Warehouse* foi realizado através do uso dessa linguagem. Assim o grupo achou por bem manter o raciocínio no processo de atualização.

Como anteriormente referido, foi criada uma área de retenção em memória com recurso a estruturas onde foi efetuado o processo de ETL. Consequentemente, foi necessário criar a mesma área de retenção com os dados já presentes no *Data Warehouse*. Para tal, foi necessário efetuar uma conexão ao SGBD que armazena o *Data Warehouse*. Para este efeito, foram utilizados os *drivers* que possibilitam tal conexão, de forma a ler e armazenar a informação nas estruturas anteriormente apresentadas através das devidas consultas.

4.6.1 Inserção de nova informação

Para inserir a nova informação, foi estabelecida a conexão à Base de Dados onde se iterou sobre todas as dimensões e factos, de forma a que as estruturas utilizadas no povoamento possuam a respetiva informação.

Posteriormente, é efetuado o mesmo processo que foi apresentado na secção 4.4. No entanto, em vez de extrair a informação do *dataset* escolhido, é extraído de um novo *dataset*, criado para o efeito que possui apenas informação para ser inserida. É de realçar que este novo ficheiro possui exatamente os mesmos campos do *dataset* selecionado.

Consequentemente, foi efetuado o mesmo processo de tratamento dos valores nulos utilizado e demonstrado anteriormente.

Encontrando-se a informação consistente, é efetuado um carregamento sequencial dos dados, tal como realizado no processo de povoamento, começando nos elementos mais simples para os mais complexos. Contudo, nesta atual aplicação, são apenas relevantes os valores diferentes dos presentes no *Data Warehouse*, ou seja, em vez de no final da aplicação percorrer as estruturas que dão suporte à área de retenção, agora após ser encontrada uma informação diferente, para além de introduzir nas estruturas tal como era efetuado anteriormente, inserimos essa informação no ficheiro final criado.

```
...
String[] date = hostSince.split("/");
String first_date = "'" + date[2] + "-" + date[1]
    + "-" + date[0] + "'";
if (!(dates.contains(first_date))) {
    String insert = "insert into dim_date values (" +
        first_date + "," + date[2] + "," + date[1] +
        "," + date[0] + ");";
        out.println(insert);
        out.flush();
}
String key_host = "" + hostID + ";" + first_date + ";"
    + hostLocation + ";" + hostResponseTime + ";"
    + hostResponseRate + ";"
    + hostTotalListingsCount + ";" + id;
if (!hosts.containsKey(key_host)) {
    hosts_counter += 1;
    Host l = new Host(hosts_counter, hostID,first_date,
    host_location.get(hostLocation),
    host_responsetime.get(hostResponseTime),
    hostResponseRate, hostTotalListingsCount);
    hosts.put(key_host, l);
    String insert = "insert into dim_host values ("
        + l.getId() + "," + l.getId_host()
        + "," + first_date + "," + l.getFk_locationhost()
        + "," + l.getFk_responsetime() + "," + l.getResponse_rate()
        + "," + l.getTotal_listings_count() + ");";
    out.println(insert);
    out.flush();
} ...
```

4.6.2 Atualização da informação

Para manter um *Data Warehouse* atualizado, foi estabelecida uma conexão à Base de Dados que dá suporte ao *Data Warehouse* de forma a iterar sobre todas as dimensões e factos de modo a que as mesmas estruturas utilizadas no povoamento possuam a respectiva informação.

Uma vez completas as estrutras, é efetuada a extração da informação a atualizar de um novo *dataset*, com a mesma informação que o *dataset* utilizado. Como o identificador da hospedagem é um facto consolidado (é o que identifica cada linha), o primeiro passo para a atualização está no tratamento dos valores nulos de cada linha do *dataset* destinado à atualização. Posteriormente, obtém-se o objeto que represente um facto através do identificador. Caso exista, é feita uma verificação dos diversos campos onde se verifica se um determinado valor está presente na estrutura respetiva. Caso contrário, é sinal que é um valor desconhecido existindo a necessidade de efetuar a adição do mesmo, bem como das respetivas dimensões. Subsequentemente, é atualizada a tabela de factos, sendo necessário verificar se o valor que está atualmente na tabela de factos corresponde ao da linha a analisar. Caso esse valor não corresponda, é então efetuado o *update* da tabela de factos com o respetivo valor. Seguidamente, é apresentado o processo de atualização da experiência oferecida. As restantes dimensões seguem o mesmo processo de raciocínio.

```
...
Facts facts = new Facts();

for (Facts f : factsArrayList) {
    if (f.getId() == Integer.parseInt(id)) {
        facts = f;
    }
}

if (facts.getId() != 0) {
...
if (!experiences.containsKey(experiencesOffered)) {
    exp_counter += 1;
    experiences.put(experiencesOffered, exp_counter);
    String insert = "insert into dim_experience values (" + exp_counter +
        ",'" + experiencesOffered + "');";
    out.println(insert);
    out.flush();
    int id_experience = experiences.get(experiencesOffered);
    String s1 = "update fact_AIRBNB set fk_experience = " + id_experience +
        " where id=" + id + " ;";
    out.println(s1);
    out.flush();
} else {
    int experience = experiences.get(experiencesOffered);
    if (experience != facts.getExperience()) {
        int id_experience = experiences.get(experiencesOffered);
        String s1 = "update fact_AIRBNB set fk_experience = " +
            id_experience +" where id=" + id + " ;";
        out.println(s1);
        out.flush();
    }
}
...
}
```

5 *Business Inteligence*

A utilização do *Tableau Software* possibilitou a visualização dos diferentes dados e análise de diversos indicadores, com o intuito de responder às questões previamente colocadas.

5.1 Análise de Resultados

Neste capítulo, são respondidas todas as questões presentes na secção 4.2 e apresentados os indicadores que possibilitaram a resposta das mesmas.

1. **Preço**

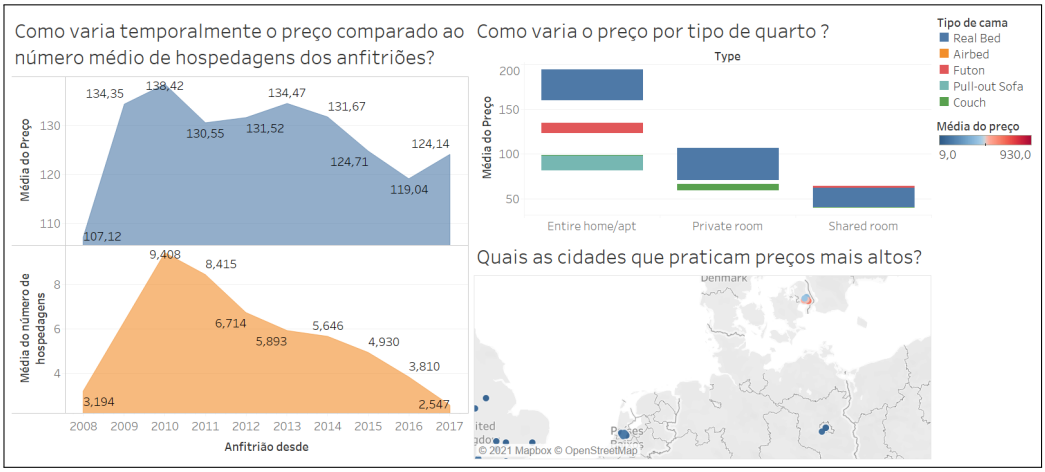

Figura 3: Resultados relativos aos preços praticados.

Os dados obtidos permitem afirmar que os preços variam consoante o tipo de quarto. Observando o gráfico (Figura 3), um quarto partilhado fica mais barato do que um quarto privativo, assim como o aluguer de um apartamento ou casa fica mais caro que um quarto privativo. De notar que um apartamento/casa apenas com um sofá cama é, por vezes, mais barato do que um quarto privativo com uma cama real.

Pela análise do mapa, verifica-se que a cidade onde são praticados os preços mais elevados é Copenhaga, sendo que o preço médio ronda os 550$.

O preço exercido pelos anfitriões altera conforme o número médio de hospedagens destes. Em 2008, o preço médio era de 108.12$ e o número de hospedagens de cada anfitrião era, em média, 3.194. Chegando a 2010, o número de hospedagens por anfitrião triplicou para 9.408 e registou-se um aumento considerável do preço (138.42$). Este aumento pode, eventualmente, dever-se a uma espécie de oligopólio, em que não existem muitos anfitriões com diversas hospedagens. Como tal, estes conseguiam praticar preços mais elevados, sendo que as alternativas seriam outro anfitrião também com muitas hospedagens. A partir de 2010, observa-se uma redução constante do número médio de hospedagens por anfitrião, explicado pelo aumento do número de anfitriões com uma única hospedagem. Verificou-se, ainda, a redução do preço médio, como resposta à concorrência entre os vários anfitriões com uma hospedagem.

2. **Preço semanal e mensal**

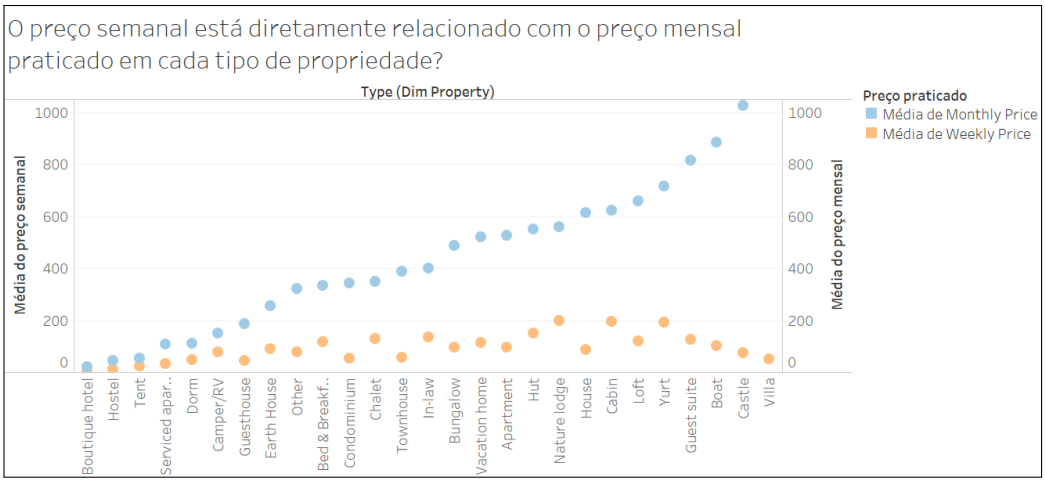

Figura 4: Resultados relativos aos preços semanais e mensais praticados.

A análise do gráfico apresentando revela uma diferença não linear entre os preços semanais e mensais praticados em cada tipo de propriedade. Assim, constata-se a existência de tipos de propriedade em que o preço mensal atinge valores dez vezes mais elevados do que o preço semanal, e outras em que o preço mensal é apenas três vezes mais elevado do que o semanal.

3. **Depósito de segurança**

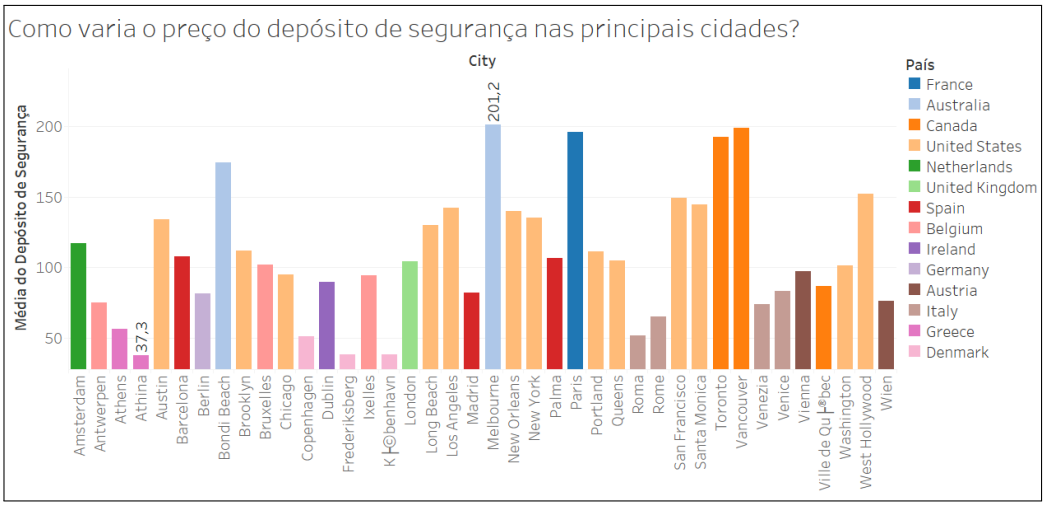

Figura 5: Resultados dos depósitos de segurança.

A representação gráfica da média dos preços de depósitos de segurança revela valores mais elevados para as cidades da América do Norte. As cidades Europeias são, em média, as que cobram menos pelo depósito de segurança, com exceção de Paris, que é uma das cidades no mundo que pratica valores mais elevados de depósito de segurança. De realçar a prática de preços elevados em algumas cidades na Austrália.

4. **Taxa de limpeza**

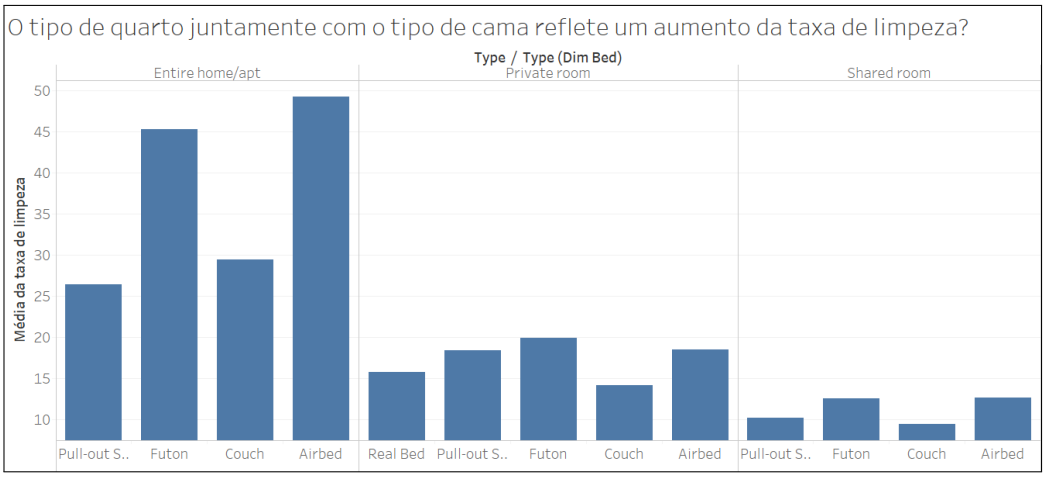

Figura 6: Análise dos resultados sobre a taxa de limpeza.

Como se pode observar pelo gráfico, a taxa de limpeza varia de forma consistente com o tipo de quarto/propriedade. Assim, os custos associados às taxas de limpeza são mais elevados em moradias/apartamentos, seguidos de quartos privativos e, por fim, quartos partilhados. O tipo de cama provoca alguma flutuação da taxa de limpeza. Nos quartos privativos e partilhados, esta taxa é negligível. No entanto, para moradias/apartamentos, o preço apresenta uma grande variância (por exemplo, um sofá-cama tem uma taxa de limpeza de metade de uma cama real).

5. **Número de hóspedes e número extra de hóspedes**

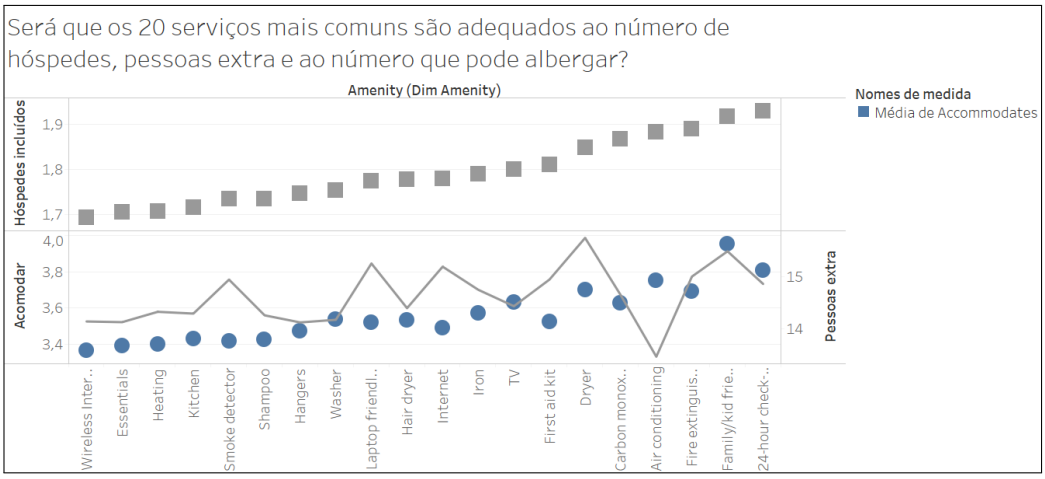

Figura 7: Resultados obtidos face ao número de hóspedes e número extra dos mesmos.

No gráfico acima apresentado, pretende-se observar de que forma os serviços mais comuns são adequados ao número médio de hóspedes que pode albergar. Para além disso, pretende-se comparar os números médios de pessoas extra existentes com os hóspedes a que determinada hospedagem é destinada.

Neste sentido, conclui-se que, em média, uma determinada hospedagem tem capacidade para acomodar três a quatro pessoas e capacidade extra de 13 a 14 pessoas. Apesar disso, a média de hóspedes incluídos é tendencialmente duas pessoas. Esta observação indica que, em média, os serviços disponibilizados são divididos igualmente, ou seja, quando um determinado serviço está vocacionado para quatro pessoas normalmente esse serviço é direcionado para dois hóspedes.

6. **Mínimo de noites**

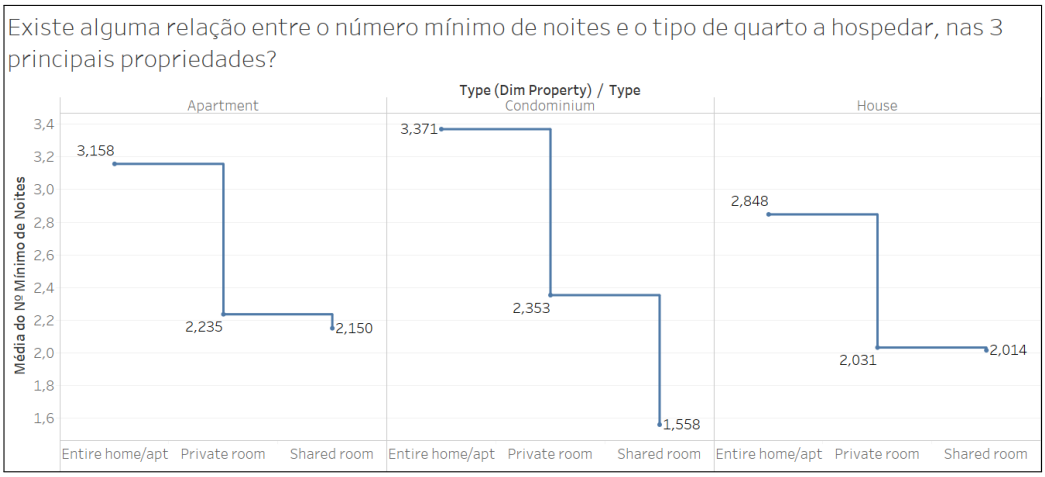

Figura 8: Resultados relativos à média de noites mínima.

O número mínimo de noites varia de forma constante pelos diferentes tipos de quartos, sendo que uma moradia/apartamento tem, em média, um número mínimo de noites superior a um quarto privativo ou partilhado.

No que diz respeito a um apartamento ou moradia, o número mínimo de noites de estadia são semelhantes. O preço praticado por quarto é mais elevado para moradias ou apartamentos, enquanto o preço para quartos privativos e partilhados são idênticos. No entanto, para uma propriedade pertencente a um condomínio, o número mínimo de noites é menor para um quarto partilhado quando em comparação com um quarto privado.

7. **Disponibilidade (anual)**

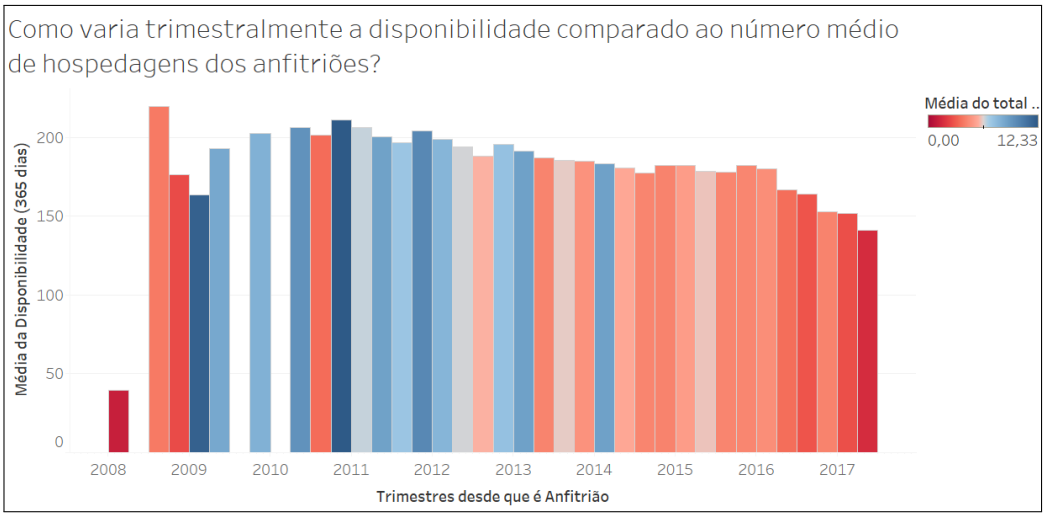

Figura 9: Dados relativos à disponibilidade anual.

O gráfico 9 pretende analisar a influência do número médio de hospedagens que cada anfitrião possui desde o seu registo na plataforma, relacionando a existência de uma disponibilidade média anual mais elevada/reduzida nesse ano. Assim, observa-se que anfitriões registados há mais tempo, e com várias hospedagens, apresentam maior disponibilidade do que anfitriões mais recentes. Neste mercado, este gráfico encontra-se relacionado com as conclusões retiradas na métrica preço, presente no gráfico 3, pelo que os anfitriões mais recentes entram no mercado com preços mais baixos, de forma a cativar mais clientes, o que se reflete numa baixa disponibilidade.

8. **Politica de Cancelamento**

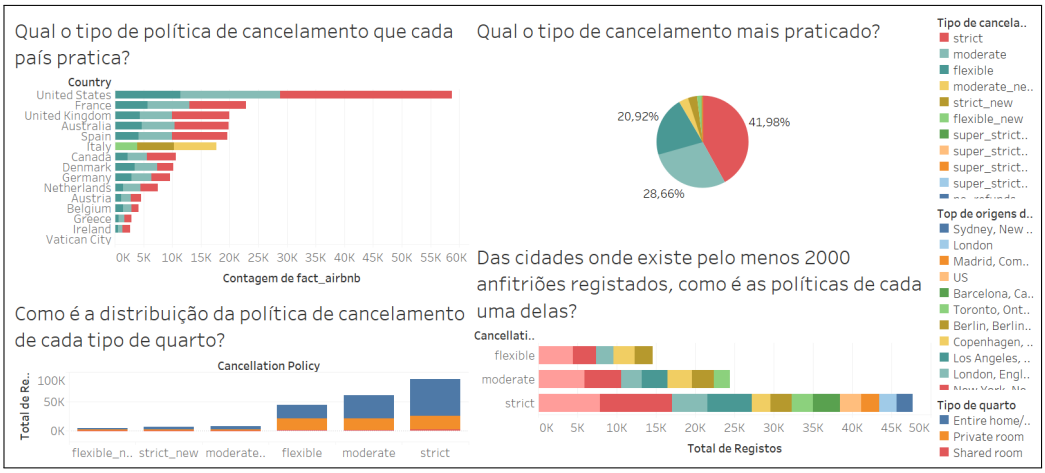

Figura 10: Dados sobre as políticas de cancelamento.

Como se pode observar, todos os países têm, em média, uma distribuição das diferentes politicas de cancelamento idênticas.

De entre as diversas hospedagens, o tipo de cancelamento mais praticado é o '*strict*' com 41.98%, seguido do '*moderate*' com 28.66% e do '*flexible*' com 20.92%. As restantes politicas são variantes das politicas de cancelamento já abordadas.

Por cidade, verifica-se uma tendência consistente de um número maior de politicas de cancelamento '*strict*' seguido por números mais reduzidos de políticas '*moderate*' e '*flexible*'. De notar que, em algumas cidades, como Madrid, existe uma politica de cancelamento quase exclusivamente '*strict*'.

Quando o tipo de quarto é uma moradia/apartamento, a politica '*strict*' é a mais utilizada pelos anfitriões. No entanto, quando o tipo de quarto é apenas privativo, não existe uma politica de cancelamento preferencial. Não foi possível obter qualquer informação relativa à politica de cancelamento de quartos partilhados.

9. **Tipo de Experiência oferecida**

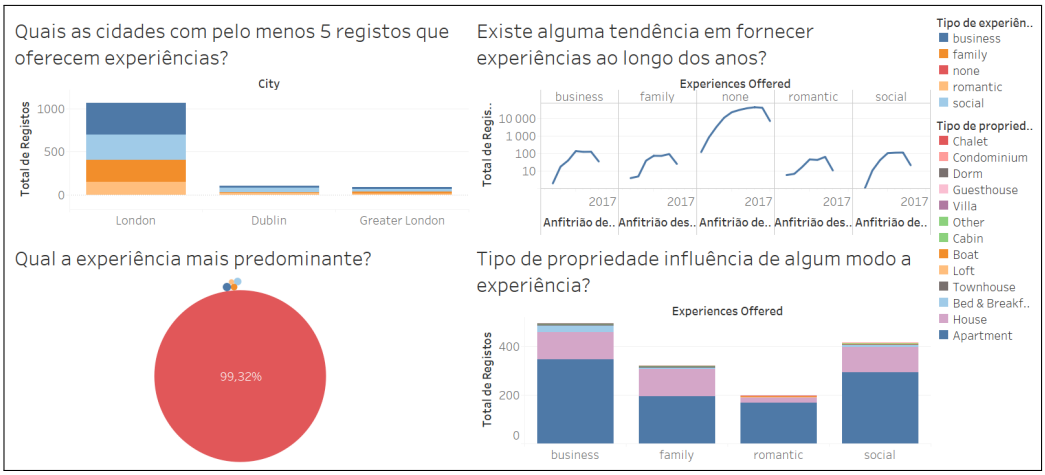

Figura 11: Dados relativos ao tipo de experiência providenciado.

No painel apresentado na Figura 11, foi efetuada uma análise da experiência que é proporcionada pelas hospedagens. Desta forma, constata-se que 99,32 % das hospedagens presentes no *dataset* não fornece qualquer experiência. Para além disso, os anfitriões mais recentes apostam cada vez mais em proporcionar novas experiências (por exemplo, anfitriões com maior tempo de registo e que proporcionam experiências fora do comum não chegam às 100 hospedagens passados quase três anos. Por outro lado, os anfitriões mais recentes registam quase 100 vezes mais hospedagens). Estas novas experiências podem ser encontradas maioritariamente em Londres e em Dublin, destinos ótimos para negócios e experiências sociais. Estas experiências fora do comum encontram-se, na sua maioria, em apartamentos e moradias.

10. **Características**

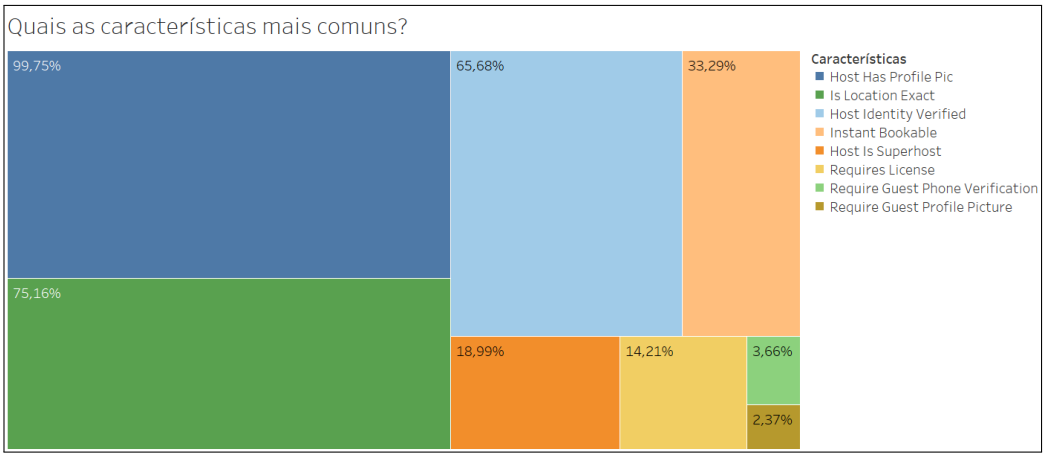

Figura 12: Dados sobre as Características.

Na figura 12, observa-se a distribuição das percentagens das diversas características que cada anúncio da plataforma possui. Como previamente mencionado, um registo pode conter diversas características, ou seja, as percentagens acima apresentadas são relativas ao total de registos. Com base no gráfico, conclui-se que quase todos os anfitriões apresentam uma imagem de perfil. Além disso, mais de metade dos casos indicam a localização exata do estabelecimento. Pela análise do indicador apresentado, é também possível concluir que, em mais de metade dos registos, o anfitrião encontra-se verificado pelo Airbnb.

11. **Outras**

- **Serviços**

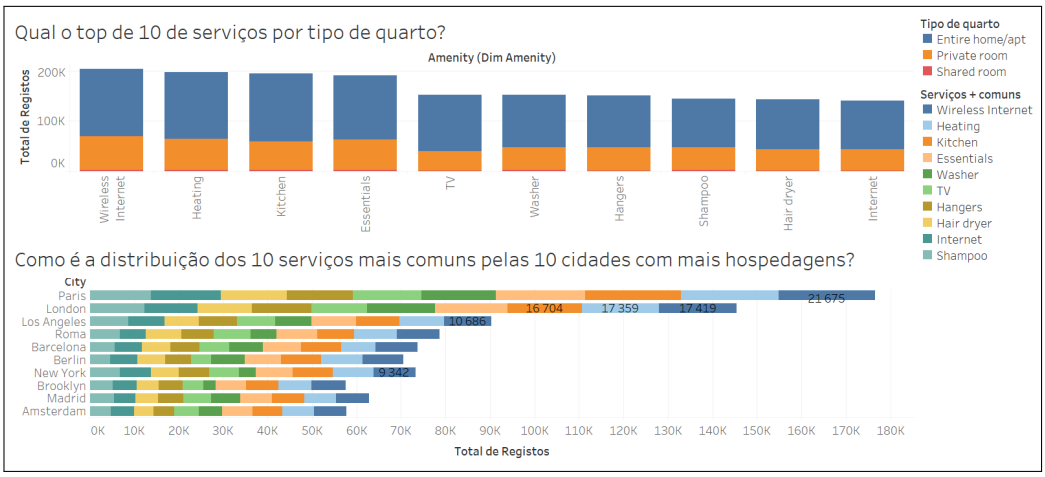

Figura 13: Serviços mais recorrentes.

Os serviços mais comuns são Internet sem fios, aquecimento e cozinha. Para tipos de quartos diferentes, os serviços mais comuns são idênticos, com exceção da televisão que, em quartos privativos, é um dos serviços menos comuns. Nas dez cidades com mais hospedagens, os dez serviços mais comuns estão presentes na esmagadora maioria delas.

- **Localização**

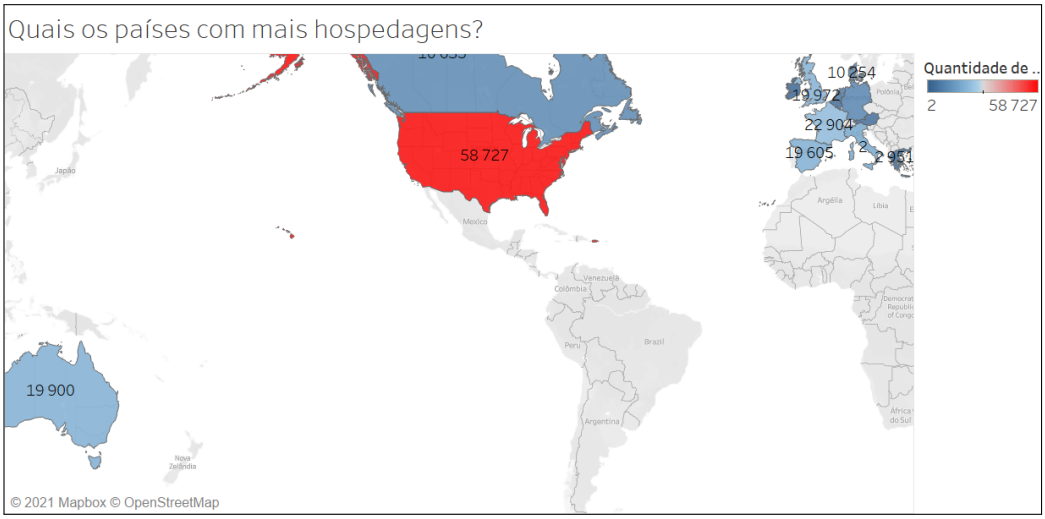

Figura 14: Distribuição das hospedagens por país.

No mapa apresentado na Figura 14, pretende-se analisar a distribuição mundial dos registos presentes no *dataset*. Desta forma, verifica-se a existência de cerca de 58727 registos associado aos Estados Unidos da América, tornando-o o país mais representado. Para além disso, Austrália, França e Reino Unido possuem bastantes propriedades, não superando, no entanto, os valores observados pelos Estados Unidados da América.

- **Anfitrião**

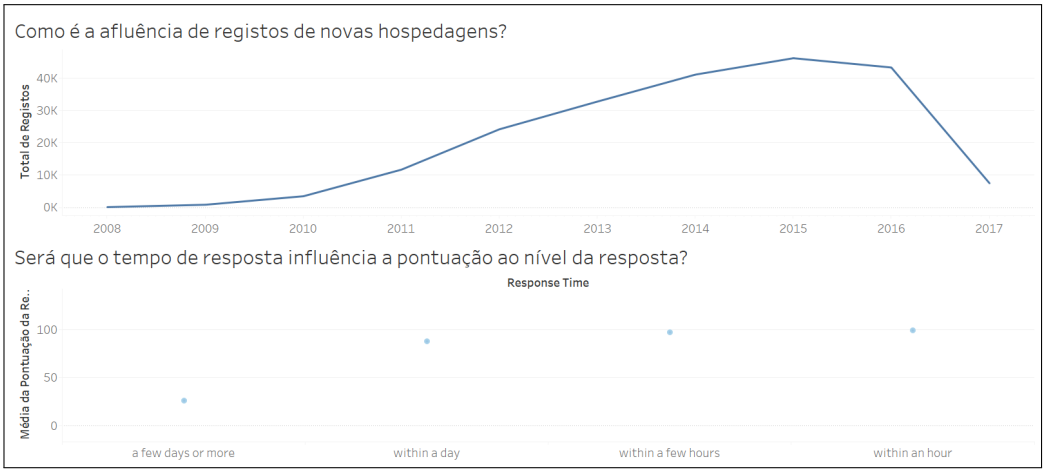

Figura 15: Propriedades da afluência dos anfitriões; Influência do tempo de resposta.

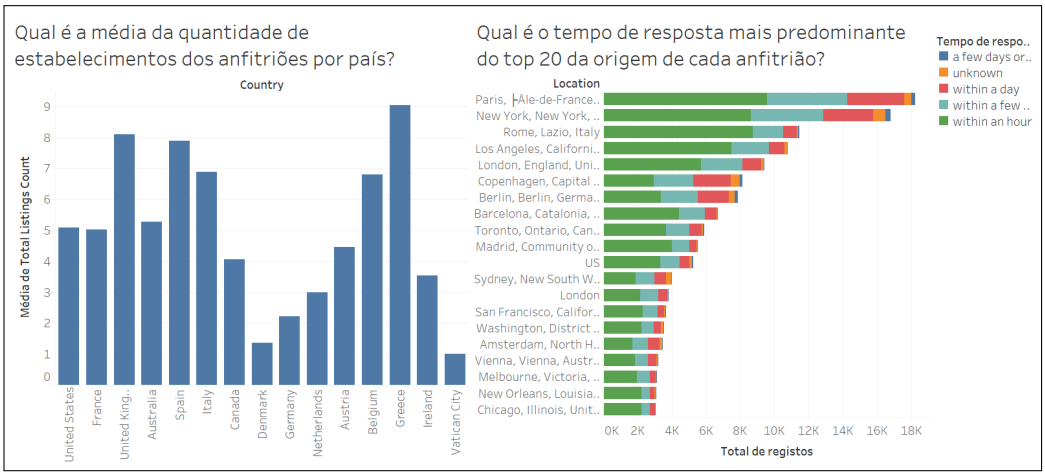

Figura 16: Origem dos anfitriões; Número de total de estabelecimentos que cada anfitrião.

Através do gráfico da Figura 15, observa-se um aumento exponencial do número de anfitriões registados num intervalo entre 2010 e 2015. Em 2010, verificavam-se menos de dez mil anfitriões registados. Cinco anos mais tarde, este valor quadruplicou. Estes resultados podem ser explicados pelo período de ganho de popularização do *Airbnb* no mercado de hospedagens *online* (uma vez que surgiu em 2008). O decréscimo observado pode dever-se à contabilização dos dados durante o decorrer do ano, pelo que este valor pode não corresponder ao do final desse mesmo ano.

A pontuação atribuída a anfitriões está diretamente relacionada com o tempo de resposta por parte destes. Assim sendo, verifica-se uma média de pontuações nos 100 valores para anfitriões cujo tempo de resposta é menor do que 24 horas. Para os restantes anfitriões, a sua pontuação será inferior a 50. O tempo de resposta não está dependente da localização do anfitrião (Figura 16, segundo gráfico). De forma geral, estes tempos de resposta são inferiores a uma hora, pelo que a percentagem de anfitriões cujo tempo é superior a esse intervalo é muito reduzida.

No primeiro gráfico (Figura 16), encontra-se a média do número de estabelecimentos dos anfitriões, por país. Deste modo, é evidente que os anfitriões possuem, em média, um estabelecimento, na Dinamarca (explicando o gráfico previamente apresentado - Figura 3, relativamente à prática de preços elevados). Por outro lado, a Grécia (país vagamente representado no *dataset*), apresenta uma média de estabelecimentos por anfitrião muito elevada. Estes dados permitem aferir que existe uma pequena comunidade que possui grande parte das propriedades neste país.

- **Verificações**

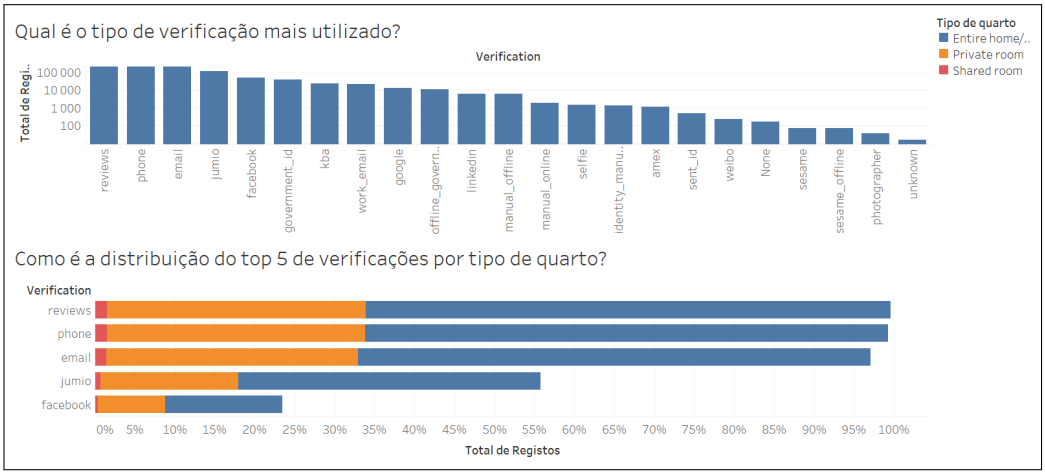

Figura 17: Dados relativos às várias verificações.

Na Figura 17, constata-se quais as verificações mais comuns. Assim, é conclui-se que grande parte dos utilizadores baseia-se nas pontuações dadas por outros utilizadores, bem como na utilização do telemóvel, email, *Facebook* e a plataforma *jumio*. A utilização destas é tem como objetivo primordial a validação da reserva de uma casa ou apartamento na totalidade.

- **Estabelecimento**

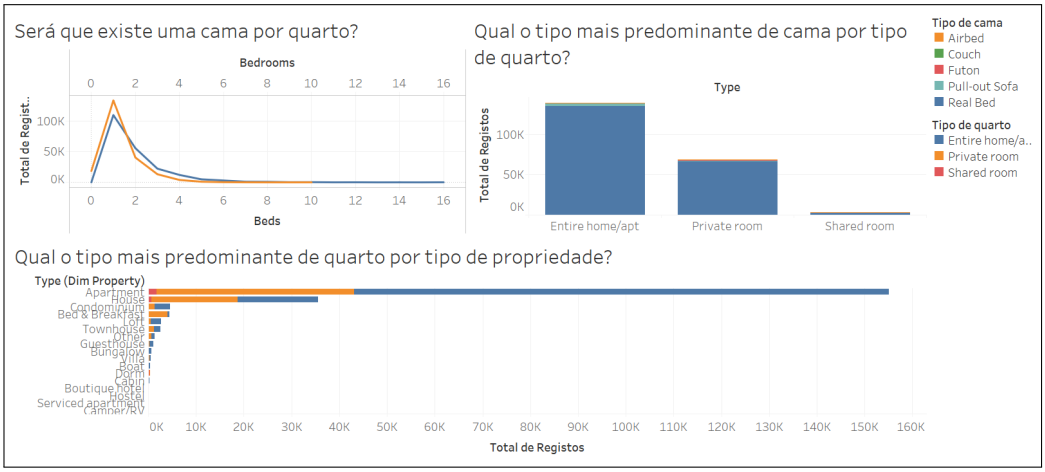

Figura 18: Propriedades dos estabelecimentos.

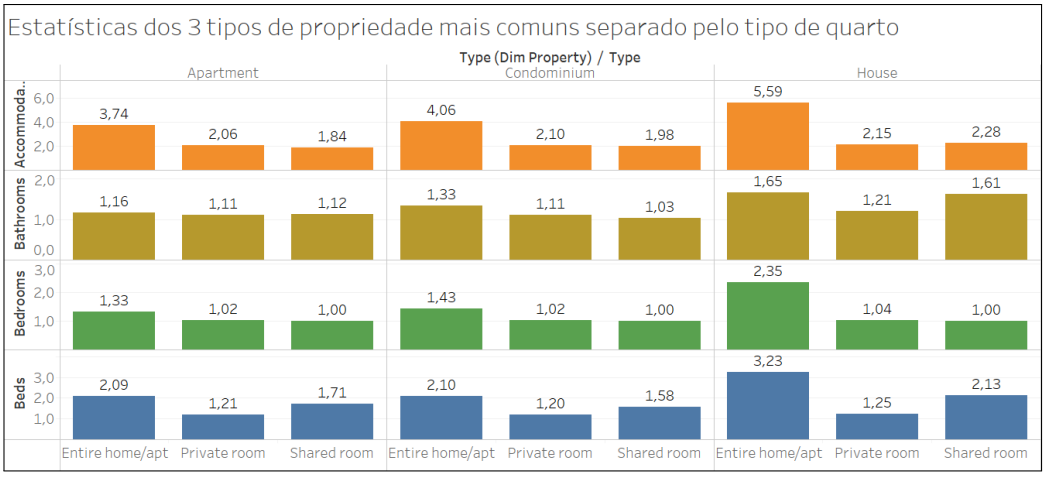

Figura 19: Propriedades das estatísticas dos estabelecimentos.

Em grande parte das hospedagens, o tipo de cama disponível é uma cama real. O número médio de quartos e camas são idênticos, como se pode observar no primeiro gráfico da Figura 18. Nos diferentes tipos de propriedades, existe alguma discrepância entre o aluguer de uma propriedade na sua totalidade e parcialmente. Em caso de apartamento, este é, na maioria das vezes, alugado na sua totalidade. Por outro lado, uma casa é igualmente alugada na sua totalidade ou apenas parcialmente. De notar que, para propriedades do tipo *Condomínio*, estas encontram-se para alugar na sua totalidade. No que diz respeito a uma propriedade do tipo *Bed and Breakfast*, tenciona-se que o aluguer seja para apenas um quarto.

Uma propriedade alugada por inteiro tem, em média, mais serviços e camas do que quartos individualmente alugados. O número de casas de banho, no entanto, apresenta uma baixa variância entre os diferentes tipos de propriedade e aluguer.

- **Pontuações**

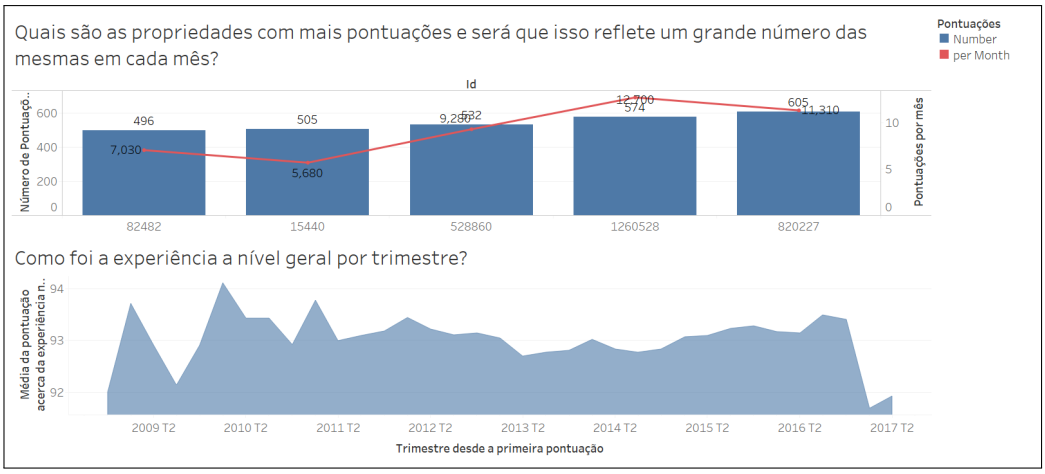

Figura 20: Dados relativos ao número de pontuações, pontuações mensais e pontuação gerais.

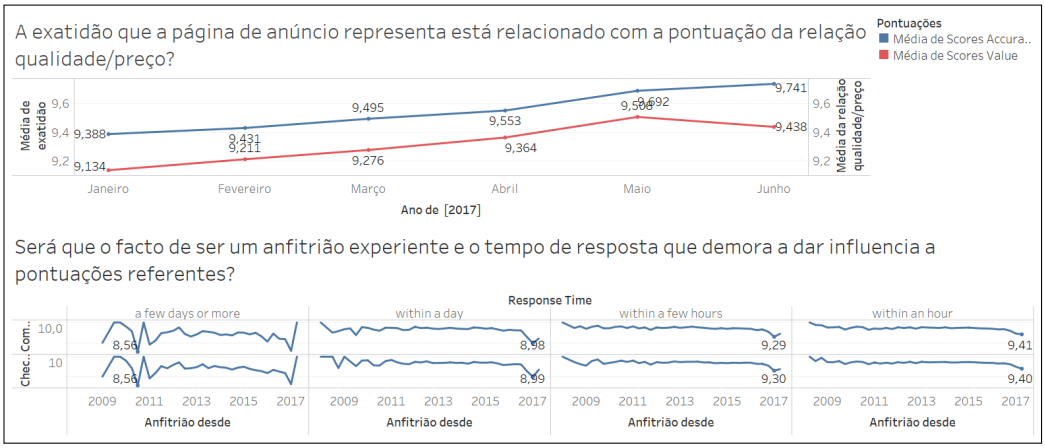

Figura 21: Relação entre a pontuação média da qualidade/preço e a precisão do anúncio; Influência da experiência do anfitrião na pontuação da comunicação e *check-in*.

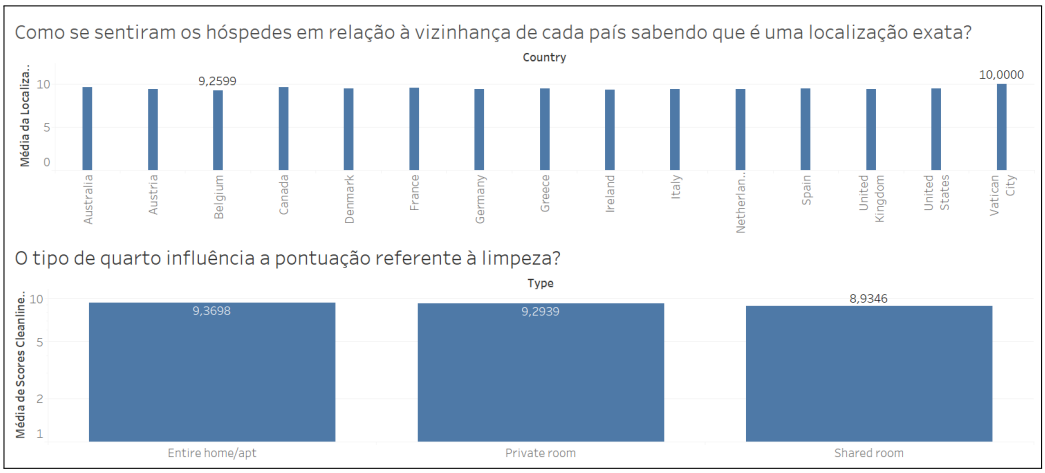

Figura 22: Dados relativos a pontuações de limpeza e localização.

Com base na Figura 21, observa-se que a média de pontuações para um tipo de quarto permanece constante, num intervalo de 92 a 94. Um quarto com uma boa qualidade/preço tem maior probabilidade de ter uma boa pontuação.

Se a página do anúncio for exata (isto é, as descrições e imagens apresentadas representam a realidade), aumenta, também, a probabilidade de receção de uma boa pontuação, como se pode observar no gráfico para o primeiro semestre de 2017.

A análise do gráfico permite corroborar a conclusão previamente inferida na Figura 15. Desta forma, confirma-se que um anfitrião com um tempo de resposta maior pode apresentar uma maior variância da pontuação recebida por parte dos utilizadores. É de salientar que os anfitriões mais recentes exibem, normalmente, pontuações mais reduzidas.

Por fim, verifica-se (Figura 22) que tanto a localização do quarto como o respetivo tipo não afetam significativamente a sua pontuação. Relativamente à pontuação de limpeza, esta tende a ser menor quando se trata de um quarto partilhado.

6 Conclusão

O presente relatório descreveu o processo de implementação de um sistema de *Data Warehousing*, bem como um sistema de *Business Intelligence*, para suporte à decisão na área de negócio de aluguer de propriedades pela plataforma **Airbnb**. Para este efeito, foi utilizado como fonte um *dataset* público.

Com a realização deste projeto, o grupo apreendeu as potencialidades que a área de *Business Analytics* pode trazer no aperfeiçoamento de diversos aspetos em qualquer área e negócio, passíveis de exploração através de dados históricos adquiridos.

Considera-se que os principais objetivos desde projeto foram cumpridos. No entanto, a escolha de um *dataset* desta extensão revelou um peso adicional no processamento do mesmo. Apesar disso, esta adversidade foi solucionada com o recurso a ferramentas e conhecimentos previamente adquiridos.

A realização deste trabalho permitiu a consolidação dos conhecimentos obtidos durante a unidade curricular, e proporcionou o desenvolvimento da capacidade de aplicação de vastas estratégias e métodos em situações reais.

Referências

[1] *Airbnb Listings*. 2021. URL: https://public.opendatasoft.com/explore/dataset/airbnb-listings/information/?disjunctive.host_verifications&disjunctive.amenities&disjunctive.features.

[2] *Inside Airbnb*. 2021. URL: http://insideairbnb.com/index.html.