Final Report - Visualização Avançada de Dados

Filipa Capela - 2018297335 - uc2018297335@student.uc.pt
Tiago Conceição - 2021167993 - uc2021167993@student.uc.pt

1 Introdução

Com este projeto pretendemos que seja exposta de forma clara, e de forma objetiva os aspetos mais relevantes dos datasets, de modo a conseguirmos retirar as conclusões mais importantes sobre estes. Assim sendo, iremos utilizar um dataset dos airbnbs (marketplace online que permite que as pessoas aluguem um quarto em determinada localização pretendida: secção 4) das grande cidades portuguesas: Porto e Lisboa. A partir daí, queremos retirar os aspetos mais importantes, comparando as duas cidades, como por exemplo: perceber quais as diferenças entre elas e qual a que se enquadra melhor num utilizador, dependendo dos objetivos que tem em consideração, isto é, o tipo de quarto que prefere, os melhores airbnbs disponíveis (posteriormente iremos abordar as questões de forma mais aprofundada). Pretendemos que os dados de visualização que obteremos alcancem um desfecho que permitam realizar uma apreciação sobre os airbnbs nas grandes cidades portuguesas e que, os utilizadores consiguam, de uma forma intuitiva, através de gráficos de visualização, escolher qual das melhores opções se enquadram no seu perfil, de forma a aumentar a sua satisfação na estadia.

Queremos abordar este tema devido à infinidade de airbnbs disponíveis, que por vezes dificulta a escolha de um para alugar. Apesar do site do airbnb [1] conter dados estatísticos, nós pretendemos comparar as duas grandes cidades e as zonas que as rodeiam, relacionando os preços com as reviews e as alturas do ano. O público alvo deste projeto, são pessoas que viajam (estrangeiros ou não) que necessitam de alugar um quarto.

Elaboramos as seguintes questões que pretendemos solucionar com este projeto:

- Quais são as zonas das cidades em que existem airbnb e verificar quais são os top X (por exemplo 10).
- Por cada cidade pretendemos verificar se existe alguma correlação entre os preços e as reviews, por exemplo, quanto mais reviews existem e houver mais opiniões positivas, se irá afetar o preço.

Assim sendo, com estas questões conseguimos entender melhor o impacto do airbnb nas diferentes cidades.

2 Related Work

No contexto deste projeto, e após alguma pesquisa, notámos que existem algumas formas de visualização que gostaríamos de integrar no trabalho, de forma a tornar a interpretação dos dados mais fácil e eficaz [3]:

- Mapa demográfico com marcação dos airbnbs disponíveis;
- Mapa demográfico com as regiões das cidades (neighbourhoods) mais requisitadas/populares;
- Gráfico circular que, nos neighbourhoods fosse possível representar a percentagem dos tipos de quarto existentes e a requisição de cada tipo de quarto durante um ano;
- Gráfico de barras que mostre para os neighbourhoods mais populares, a média de preço por noite para cada tipo de quarto;
- Gráfico de linhas que mostre a percentagem de ocupação do melhor quarto nos melhores neighbourhoods de cada cidade.

3 Work Plan

Para este trabalho definimos um certo calendário para as tarefas serem desenvolvidos nos prazos para não haver trabalho acumulado.

Tarefas	Prazos
Procura do dataset e Preparação do Paper	27-02-2022
Escrita do Paper e Inicio do Pré-processamento	06-03-2022
Fim do pré-processamento Começo da Exploração de análise de dados	13-03-2022
Continuação da Exploração de análise de dados Início do desenvolvimento dos mockups	20-03-2022
Conclusão da Exploração de análise de dados Finalização dos mockups Desenvolvimento do report/powerpoint da apresentacão intermédia	07-04-2022
Estabelecimento das melhores estratégias de design	09-04-2022
Inicio da implementação dos dados com suporte em D3	18-04-2022
Preparação da apresentação e entrega final	19-05-2022

4 Data

Numa primeira análise, foi selecionado o dataset de Airbnb de Lisboa, mas como em Portugal só existem registos das grandes cidades (Lisboa e Porto), achámos que seria uma mais valia usar ambos, para uma melhor base de comparação entre os dois. Estes datasets contêm as mesmas características pelo que facilita a sua análise, e o simples facto de estarem bastante completos a nível de detalhes fornecidos, torna a sua visualização mais eficaz. Cada dataset é composto por vários ficheiros, no entanto iremos apenas utilizar : *calendar.csv*, 2 ficheiros *listings.csv*, *neighbourhoods.geojson*, *neighbourhoods.csv* Com os dados fornecidos dá para fazer algumas análises interessantes, que podem ser uma mais valia para quem consulta esta informação. Estes datasets são originais do

próprio site do airbnb pelo que todos os dados são fidígnos e estão de acordo com a realidade [2]. É importante referir que estes datasets foram os fornecidos pelos docentes no enunciado do projeto.

5 Exploratary Data Analysis

Para iniciar a análise exploratória de dados foi necessário realizar o pré-processamento das tabelas do dataset. Desta forma foi fundamental descartar algumas tabelas desnecessárias, remover colunas e juntar tabelas (no dataframe do calendário foi adicionada uma nova coluna com o neighbourdhood-group associado ao id do quarto), para que seja possível incorporar a informação disponível de todas tabelas para a análise de dados. Posteriormente foi retirado todas as linhas com valores inválidos, pois não era possível realizar nenhuma ação com eles ou seja, tentar gerar um valor aleatório, pois futuramente poderia interferir com a análise final dos dados (como por exemplo arranjar uma média de preço/review é assim corresponderia à realidade).

Numa fase inicial, para uma análise mais geral, para cada cidade (Lisboa e Porto), decidimos analisar qual a quantidade de airbnbs disponíveis na zona e o preço médio dos quartos.

Após a análise dos gráficos obtidos, observamos que a região de Lisboa tem um número maior de alojamentos e consequentemente um preço médio mais elevado.

Para cada cidade, analisámos a distribuição dos tipos de quarto num mapa demográfico:

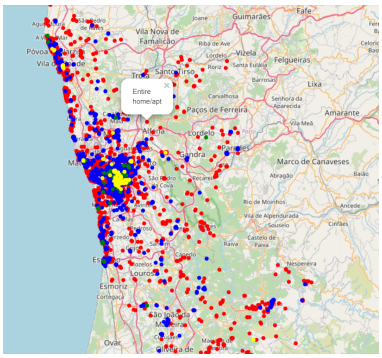

Figure 1: Localização dos tipos de quarto na região do Porto

Na figura 1 os pontos de diferentes cores representam diferentes tipos de quarto: vermelho corresponde a Entire home/apt, azul a Private room, verde a Shared room e amarelo a Hotel. Podemos observar que a maior parte dos tipos de quarto se localizam no centro do Porto, dado que é a zona mais propensa a se visitar devido aos pontos históricos.

Ao ser feita a análise dos gráficos que remetem para a média/percentagem dos tipos de quartos no porto, verificamos que os apartamentos inteiros para alugar têm um preço muito maior que alugar um quarto num hotel, tal deve-se ao facto de haver uma quantidade maior de quartos para alugar (*Entire home/apt:* cerca de 84%) do que quartos de hotéis (*Hotel room:* cerca de 0.64%), visto que é algo com uma maior procura.

De seguida foi necessário aprofundar e apresentar de forma mais detalhada cada região de modo a obter dados relevantes. Assim sendo, investigámos cada município vizinho de cada região mencionada anteriormente, ao qual resultou em quatro considerações distintas:

- Média de Preços por neighbourdhood_group;
- Média/Mediana das Reviews por neighbourhood_group.
- Percentagem de ocupação por neighbourhood_group.

Com a visualização do gráfico da média de preços por neighbourhood group, observamos que a região (neighbourhood_group) *Trofa* é que apresenta um maior preço médio comparativamente aos restantes neighbourdhoods. O gráfico da média de preços durante o ano, mantém valores de preços constantes ao longo do ano, à excepção de *Dezembro 2022*, que provavelmente se deve à Passagem do em que há uma maior procura.

Algo que se pode retirar da analise de reviews, verificamos que as review dadas são mais próximas de 5 estrelas, ou seja, há poucos utilizadores que classifição os airbnbs com valores baixos, correspondentes a má avaliações.

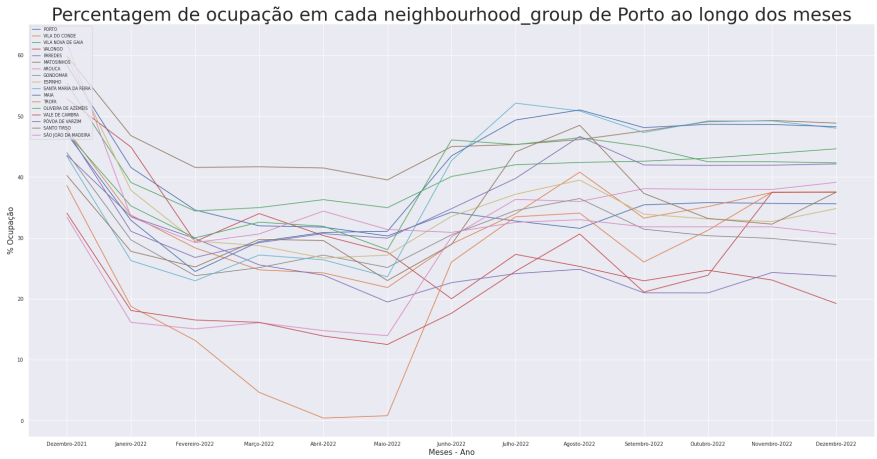

Figure 2: Percentagem de ocupação por Neighbourdhood-group durante 1 ano

Com a análise da figura 2 verificamos que apesar de o dataset ter sido disponibilizado no início de 2022, já existem airbnbs reservados para o final do ano de 2022. Assim sendo, a % de ocupação vai alterar ao longo do ano, uma vez que os utilizadores vão alugar airnbs mais próximo da data que necessitam, caso haja disponibilidade dos quartos.

De modo a conseguir perceber se existe uma correlação entre preço e as reviews, foi feito também um gráfico para cada grande cidade para ver se efetivamente existia uma correlação entre ambos os fatores. Ao ser analisada mais em detalhe vemos que de fato não existe propriamente uma relação entre os dois, porque concluímos que não é propriamente necessário ter um alojamento caro, para ter boas reviews, tal deve se à opinião subjetiva dos utilizadores e da sua expectativa perante o quarto que alugaram.

Outro gráfico ao qual retiramos informação, é o heatmap, que faz a correlação entre todas as colunas do dataset, conseguindo desta forma chegar à conclusão quais as relações entre as diferentes colunas.Assim sendo concluímo que os diferentes tipos de reviews estão fortemente relacionados com a review final.

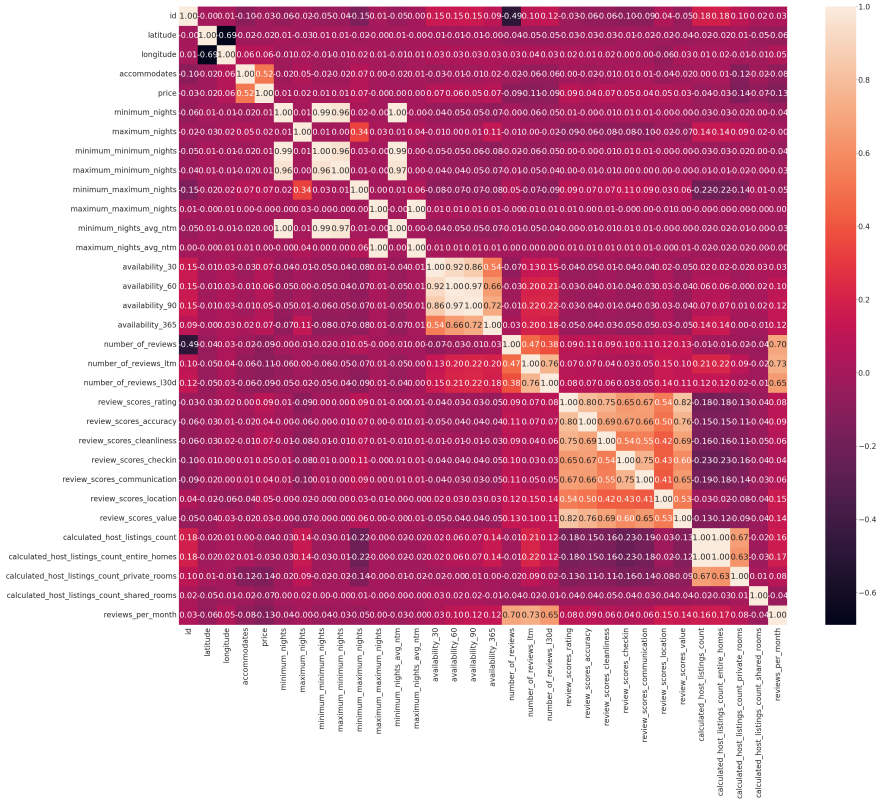

Figure 3: Heat map do Porto

6 Design

Os designs que foram idealizados para o nosso projeto, foram pensados de forma a que se conseguisse retirar o máximo de informação possível, e com um estilo apelativo. Desta forma, temos alguns gráficos básicos de barras, e circulares que vão indicar dados que não dão para agregar a outros gráficos, como por exemplo o número de alojamentos que existem tanto em Lisboa e Porto.

Porém, existem dados que exigem gráficos mais complexos, como por exemplo, a linha da média dos preços/reviews geral de todas as cidades vizinhas. Assim iremos filtrar uma cidade em concreto, para desta forma vermos o quão essa cidade está desfasada das outras.

Teremos também um gráfico mais específico em que irá conter um mapa demográfico com a média das reviews/preços dos neighbourdhoods de cada neighbourdhood_group com um gráfico ao lado com as médias das reviews/preços de cada (figura 4). Com este gráfico conseguiremos ver o comportamento dos neighbourdhoods e o que isso implicará no seu respetivo neighbourdhood_group.

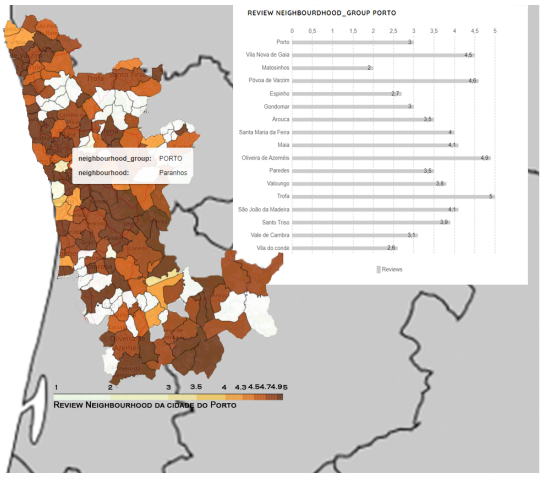

Figure 4: Média das reviews em cada neighbourhood e neighbourhood$_group$

Iremos também apresentar no site a possibilidade de mostrar os top 10 melhores quartos num determinado neighbourdhood_group ou na cidade, dependendo da escolha do utilizador. Nesta opção, iremos considerar as reviews do quarto e o preço de mesmo (figura 5).

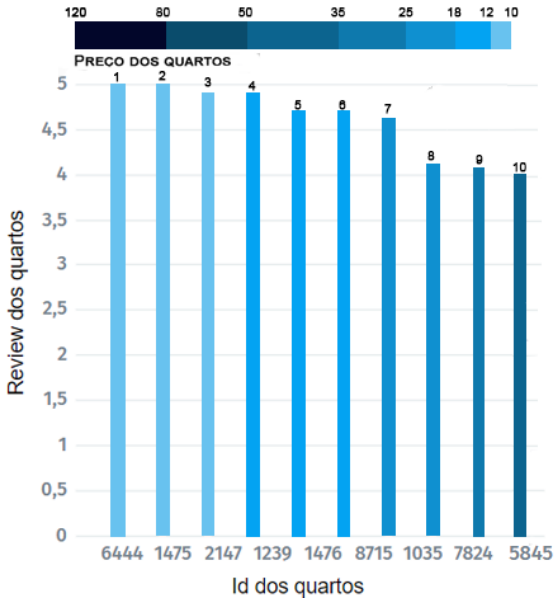

Figure 5: Top 10 melhores quartos num neighbourdhood_group do Porto

7 Implementação

Ao ser efetuada a implementação do website neste projeto, quisemos que a sua elaboração fosse o mais minimalista possível, de modo, a manter o foco nos elementos feitos em d3, concretamente os mapas e gráficos.

Desta forma, contamos com poucos elementos, na página principal, apenas temos uma barra superior, que tem como finalidade identificar o projeto.Nessa mesma página temos quatro botões

distintos que servem para encaminhar para outras páginas secundárias com outro tipo de informação de acordo com a região escolhida. Esses botões estão por baixo de cada mapa, ambos com um propósito diferente, tal como:

- Mais informações Lisboa;
- Top cinco quartos mais baratos Lisboa.
- Mais informações Porto.
- Top cinco quartos mais baratos Porto.

Assim, nas páginas secundárias, é necessário um botão para voltar, para a página principal, este vai-se encontrar no topo.

Na página dos top 5, para obter os melhores regiões demos um peso de 60% à review e 40% ao preço.

8 Avaliação

De modo a obter feedback do projeto desenvolvido, questionámos a várias pessoas a suas opinião e todas foram opiniões positivas. Apesar de o website não fornecer muito dinamismo, relativamente a opções para escolhas de parâmetros ou filtros, o site fornece uma compreensão simples e um alcance dos objetivos pré-definidos.

9 Reflexão

Com a conclusão deste projeto, podemos retirar várias reflexões, nomeadamente que a visualização dos dados pretendidos foi bem sucedida, e que apesar de não ter tanto dinamismo nas visualizações, consegue-se retirar conclusões concisas dos dados. Relativamente ás questões colocadas, conseguimos responder-lhes parcialmente através da visualização dos dados, mas no entanto com os dados recolhidos, consegue-se atender a muitas outras perguntas que não foram definidas. No entanto, com os dados deste projeto dá para aprender várias aspetos que sem elas não seria possivel, nomeadamente:

- Quantidade de quartos por região;
- Distribuição do tipo de quartos por região.
- Mapa da distribuição dos preços por região.
- Gráfico da distribuição dos preços pelos diferentes tipos de quarto em determinada região.
- Mapa da distribuição das reviews por região.
- Gráfico da distribuição das reviews pelos diferentes tipos de quarto em determinada região.
- Gráfico da variação do preço ao longo do ano.
- Mapa do top quartos mais baratos segundo uma métrica definida por região.

A ideia para este projeto, era ambiciosa mas a falta de experiência em d3, revelou-se um obstáculo, mas no entanto em futuros projetos que envolvam esta linguagem, provavelmente o resultado será melhor.

De seguida, apresentamoso link do video demonstrativo do website: https://youtu.be/1rw8_SNBUQs

References

[1] Airbnb lisbon stats. http://insideairbnb.com/lisbon/.

[2] Get the data - inside airbnb. adding data to the debate. http://insideairbnb.com/get-the-data.html.

[3] Visualização de gráficos. https://infogram.com/pt/pagina/escolha-grafico-de-visualizacoes-certo.