

**Trabalho Prático 1**

Diogo M. Riscado, Nº 51869

Tiago Miguel Fernandes Marques, Nº 51653

**Resumo:**

O relatório em questão debruçar-se-á sobre a Recolha, Integração, Análise Exploratória e Limpeza e Preprocessamento de Dados, tendo como base *databases* obtidas na plataforma *Kaggle*. Através da união das mesmas, torna-se possível comparar diferentes características entre si, tendo em conta o número de quartos, a região geográfica, a área, entre outros.

**Índice**

1. Recolha de Dados
2. Integração de Dados
3. Análise Exploratória
4. Limpeza e Preprocessamento

Conclusão Geral

1. Recolha de Dados

Como referido anteriormente, a recolha de dados cingiu-se, de certa forma, à obtenção das *datasets* através do [Kaggle](https://www.kaggle.com/datasets). É importante, porém, referir uma das primeiras decisões tomadas conjuntamente: a localização relativa dos dados. Foram escolhidos três ficheiros, cujas informações nos permitiriam retirar conclusões acerca da cidade de *Kansas City* (*Missouri/MO*, doravante *KC*), *Perth* e *Melbourne*, sendo as últimas duas localizadas em território australiano, e não nos Estados Unidos da América. Ainda assim, estas duas cidades distam 2721 KM por ar ou cerca de 3500 metros por terra entre si. Poderá ser útil compreender que *KC* encontra-se numa região relativamente central dos EUA, enquanto ambas as cidades australianas são cidades costeiras.

Assim, após esta seleção ponderada, extraímos os ficheiros .CSV e abrimo-los utilizando o *Pycharm*. De uma forma geral, o trabalho organizou-se da seguinte forma:

EIACD

│

├── datasetfiles

│ ├── kansas\_housing.csv

│ ├── melbourne\_housing.csv

│ └── perth\_housing.csv

│

├── main.py

├── recolha\_integracao.py

├── sample.py

├── README.md

└── TP1\_relatorio.docx

1. Integração de Dados

Antes de procurarmos unir os documentos criados, decidimos criar igualdades entre os .CSVs anteriores. Em primeiro lugar, excluímos variáveis que não considerámos úteis para análise, fosse por défice de informação, ou por não ser possível efetuar uma comparação passível de um juízo pertinente.

Serve de exemplo um parâmetro aplicado apenas a casas no *dataset* relativo a *Kansas City* denominado de *grade* (ou “nota”, numa tradução solta). Esta característica atribuía uma nota à casa de modo a avaliá-la comparativamente às restantes. Decidimos excluir este índice por se basear em parâmetros como “design” ou “*building construction*”, parecendo-nos importante sermos capazes de efetuar essa mesma avaliação através dos restantes dados e entre comparações dos edifícios em questão. De seguida, modificámos o nome dado às características de modo a que, ao efetuar um *join*, qualquer que fosse, houvesse maior facilidade na integração e preparação dos dados para posterior análise. Deste modo, restava-nos alterar o tipo de dado utilizado para armazenar datas- por exemplo, datas de venda- e fazer reduções minutas a poucas outras características.