**Introduction**

This notebook records my process of exploring and analyzing the [Melbourne Housing Market dataset](https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market/data). The project was inspired by Macquarie University’s *Data Visualization in Excel* course on [Coursera](https://www.coursera.org/learn/excel-data-visualization/home/info), and follows a similar—but not identical—dataset to apply and extend the techniques I learned there.

**Importing data**

Get data from csv. file and transform the data in Power Query

**Cleaning data**

* Because I want to analyze the price trend, I filtered Method to only keep “Sold” (S), “Sold after Auction” (SA) and “Sold prior to Auction” (SP) because they are valid, real price that resulted in sale.
* Cleaned and Trimmed text in Power Query.
* Remove “City Council” from the Council Area for a cleaner look.
* Change Date type
  + I encountered an error with Date Format, where Power Query cannot parse the input at date value. So, I tried to inspect and saw that PQ cannot tell if the date format was dd/mm/yyyy (4/1/2017) or mm/dd/yyyy (30/5/2017)
  + First attempt: Adjusted the Locale to United Kingdom and checked again. Problem was unresolved.

A screenshot of a computer

AI-generated content may be incorrect.



* + Second attempt: Added Column from Examples. I typed the data format that I wanted (mm/dd/yyyy) but it still didn’t solve the problem.
  + Third attempts: Changed the regional setting from US to UK. Problem solved.

A screenshot of a computer

AI-generated content may be incorrect.

* Check for duplicates
  + Inspect duplicates by grouping them. I grouped them into Address, Suburb, Date, Price and Seller Agent, and counted the row. Those with 2 rows indicated that they are true duplicates. Reasons behind the grouping:
    - Address can be identical in different Suburbs, so Address alone would not be enough to identify duplicates.
    - Each property can be sold multiple times with the same Price and same Seller Agent; hence, if Date is also the same, it’s likely duplicate.

A screenshot of a computer

AI-generated content may be incorrect.

* + Remove duplicates.
    - Count rows before removing duplicates: 43395
    - Count rows before removing duplicates: 43388 (there are 7 duplicates)
* Remove all null value for Price.

After cleaning, there are a total of **37464** rows to explore.

**Exploring**

* Extract Year and Month from date for analysis.
* When looking at the data, I saw that there were missing values for March 2016 and January 2017. This might be due to the filter I did for Method, so I went back to the original data to double-check. However, there was indeed no data for these months. This did not affect the analysis so I ignored them for now.

A screenshot of a spreadsheet

AI-generated content may be incorrect.

* Explore

|  |  |
| --- | --- |
|  | Descriptive observation |
| Melbourne’s price trend for each type over 3 years | Houses remained the most expensive, peaking around March–June before easing slightly mid-year.  Townhouses followed similar seasonal movements, while units stayed the most affordable and steady.  Overall, price trend reflect seasonal peaks in autumn (Mar-May) and spring (Sept-Nov). |
| The number of property bought over the year (Question: Are there more property bought when price is lowest or highest?) | House sales dominated the market, peaking around May and again in September–October.  Units and townhouses followed similar but smaller trends, with steadier volumes across the year.  The market is also active by seasons. |
| Top 5 suburb with the highest sales in each Council Area | Using the dashboard, user can explore in that period, which Council Area is the most active in sale performance. From there, they can dive down to the suburb’s average price as well as the top five suburb with the highest price. |
| Council Area’s average price by month and year |
| Suburb’s average price by month and year |
| Average price for each type by suburb, compared with the same month last year |