



TDS2101 Introduction to Data Science

Trimester 3, 2023/2024

Assignment Part B

Lecture Section: TC1L

Tutorial Section: TT1L

Group 4

Tutorial Tutor: Pang Nyuk Khee @ Angeline Pang

Group Members:

NAME	STUDENT ID
MUHAMMAD DHIYAU NAUFAL BIN ZAINUDDIN	1201201537
DANIEL IMTIYAZ BIN FAISAL	1201201743
NUR AYU AMIRA BINTI IDRIS	1201200722

Table of Contents

1. Questions	3
2. Data Collection	6
2.1 Primary Dataset	6
a. Dataset Description	6
b. Preliminary Insights	7
c. Describing the Dataset	7
2.2 Supplementary Dataset	8
d. Dataset Description	8
e. Preliminary Insights	8
f. Describing the Dataset	9
3. Data Pre-Processing	10
4. Data Pre-Processing for Supplementary Dataset	16
5. Exploratory Data Analysis	21
1. Summary Statistics	21
2. Question	24
2.1 Descriptive Question	24
3.2 Causal Question	40
3.3 Predictive Question	44
3.4 Exploratory Question	48
3.5 Mechanistic Question	51
6. Challenges Encountered and Future Proposal	54
7. References	55

1. Questions

Type of question: Descriptive Question

A descriptive question aims to summarise a feature of a collection of information. This kind of question seeks to understand and describe the characteristics, patterns, and relationships within a dataset. In the preliminary stages of data analysis, these questions are fundamental and will aid in gaining insight into the data being examined. The objective of descriptive questions is not to make predictions or inferences about causes. They focus instead on summarising and explaining the available data.

- **“What is the trend/inflation rate of property price for condominiums in (area / district) for the last 5/10/15 years? Plot a graph for the trend for overall condominium or only 10 property schemes with the most transactions.”**

Refined Descriptive Questions:

- **“What is the trend of property price for condominiums in the Mukim Batu for the last 5 years of 2017 until 2021? Plot a graph for the trend for overall condominium in the Mukim Batu.”**

The question was redefined from a broader question about property price trends and inflation rates over multiple time periods and locations to a more specific query focusing on the Mukim Batu area for the last five years (2017 to 2021) and asking a graph of the overall condominium trend in that area. The proposed modification enhances the level of specificity, optimises the process of gathering data, provides a clearer indication of the time frame under consideration, and specifies the requirement for visual representation. As a result, the research objective becomes clearer and more feasible to accomplish.

- **What is the median property price for condominiums in the Mukim Batu from 2017 until 2021?**

This question seeks to analyse and comprehend the median property prices in the Mukim Batu region during the specified time period (2017 to 2022). The median property price is useful because it provides insight into the market's middle price point. By focusing on a five-year period, it is possible to examine trends and price fluctuations over time. Real estate professionals, investors, and policymakers can use this information to better comprehend the market dynamics in the Mukim Batu region during this time period.

- **“What is the average/median price per square for each property scheme/condominium?”**

Refined Descriptive Questions:

- **“What is the median price per metre for each condominium scheme in the Mukim Batu from 2017 until 2021?”**

The question has been modified to enhance clarity and specificity. It was originally posed as "What is the average/median price per square for each property scheme/condominium?" and has been revised to "What is the median price per metre for each condominium scheme in the Mukim Batu from 2017 until 2021?" The revised version aims to enhance clarity by narrowing the focus to condominiums and their associated schemes. It explicitly states the objective of determining the median price for each unit size within a condominium scheme, thereby ensuring the reliability and precision of data analysis.

- **“What is the average/median price per square for each property scheme/condominium?”**

Refined Descriptive Questions:

- **“What is the average Price Per Meter for each condominium scheme in the Mukim Batu from 2017 until 2021?”**

The question was reformulated from a complex inquiry involving both average and median calculations as well as a reference to "per square" to a simpler and clearer inquiry centred on determining the average price per unit size for each condominium scheme. This modification eliminates potential ambiguity, maintains specificity by focusing on condominiums, and ensures terminology consistency throughout, thereby making the research objective more precise and direct.

Type of question: Causal Question

A causal question asks about whether changing one factor will change another factor, on average, in a population. Sometimes the underlying design of the data collection, by default, allows for the question that you ask to be causal.

“Does lot size influence the price of a condominium?”

The question "Does lot size influence the price of a condominium?" seeks to establish whether or not the size of a condominium's lot has any bearing on the asking price of a condominium. It's an attempt to determine whether or not the lot size of a condo has any bearing on its selling price. This question requires us to examine the CSV file, specifically the columns labelled "LotSize" and "Price," as well as any other elements that might affect the cost of a condo. Condominium lot size may be a causal component in price changes, and

this association can be evaluated by data examination and statistical studies or regression modelling.

Type of question: Predictive Question

A predictive question would be one where you ask what is the set of predictors or factors for a particular behaviour.

“Which schemes will have the highest value after the next 5 years of 2021?”

Using linear regression machine learning techniques, we aim to forecast which schemes will have the highest value for the next five years after 2021. This requires accumulating relevant data, cleaning and preparing it, training and testing the models.

Type of question: Exploratory Question

“What insights can be gained regarding the relationship between the size of a property and its price and Price Per Meter? Additionally, based on these relationships, could you identify the top 5 condominiums that exhibit the most favourable value?”

The examination indicates that property size has a substantial impact on both the total price and the cost-effectiveness. Bigger properties often come with higher prices but lower PricePerMeter ratios, potentially appealing to those in search of more space without substantially increased costs per square unit. Nonetheless, it's essential to factor in other variables like location and property condition when deciding on a real estate investment.

Type of question: Mechanistic Question

“How do the condominium’s location, nearby attraction, and rental management practices all come together to affect how often it is rented and how much rental income generates for 1000 square feet (sqft)”

Using additional Dataset from iproperty and property guru which is supplementary dataset we compare the prize for each scheme in one mukim and do an analysis for the question.

2. Data Collection

In this section, we will describe the datasets used in our Data Science project to analyse and visualise condominium data in Kuala Lumpur's Mukim Batu district. This section will discuss the primary datasets, their contents, preliminary findings, and any additional datasets used in the analysis.

2.1 Primary Dataset

a. Dataset Description

This project's primary dataset was obtained from the National Property Information Centre (NAPIC) website. It contains a wealth of data regarding property types such as Flat, Terrace, and Condominium in the Kuala Lumpur region over the past five years, from 2017 to 2021.

Due to the vast amount of data in the provided dataset, we initially encountered difficulties in analysing it. The dataset contains a substantial amount of 33,766 rows and 13 columns. Further exploration and data preprocessing were required to gain a better understanding of the dataset. The diagram below illustrates the structure and content of the dataset that was provided.

	STATE	DISTRICT	DateOfValuation	LotSize	PropertyType	SECTOR	PRO_TYPE	NoOffFloors	ADDRESS	SCHEME	SYER1	PRICE	YEAR
0	Kuala Lumpur	Kuala Lumpur Town Centre	27/9/2017	65	Condominium/Apartment	Residential	Condominium/Apartment	21	B-8-13,JALAN PUDU ULU	163 SERVICE SUITES	1	375000	2017
1	Kuala Lumpur	Kuala Lumpur Town Centre	28/6/2017	103	Condominium/Apartment	Residential	Condominium/Apartment	21	D-7-3A,JALAN PUDU ULU	163 SERVICE SUITES	1	503000	2017
2	Kuala Lumpur	Kuala Lumpur Town Centre	30/9/2019	104	Condominium/Apartment	Residential	Condominium/Apartment	12	2-10-3,JLN STONOR	1A STONOR APT (CONLAY COURT)	1	660000	2019
3	Kuala Lumpur	Kuala Lumpur Town Centre	1/1/2018	79	Condominium/Apartment	Residential	Condominium/Apartment	8	1-3-5,JLN STONOR	1A STONOR APT (CONLAY COURT)	1	680000	2018
4	Kuala Lumpur	Kuala Lumpur Town Centre	12/8/2018	88	Condominium/Apartment	Residential	Condominium/Apartment	12	1-1-4,JLN STONOR	1A STONOR APT (CONLAY COURT)	1	980000	2018
...
33761	Kuala Lumpur	Mukim Ulu Kelang	25/1/2017	139	Terraced House	Residential	2 - 2 1/2 Storey Terraced	2	75,JLN WANGSA BUDI 7	WANGSA MELAWATI	1	750000	2017
33762	Kuala Lumpur	Mukim Ulu Kelang	7/4/2017	108	Terraced House	Residential	2 - 2 1/2 Storey Terraced	2	4,JALAN 1/24H	WANGSA MELAWATI	1	750000	2017
33763	Kuala Lumpur	Mukim Ulu Kelang	27/8/2019	112	Terraced House	Residential	2 - 2 1/2 Storey Terraced	2	NO. 41,ORONG WANGSA SIAGA 6	WANGSA MELAWATI	1	750000	2019
33764	Kuala Lumpur	Mukim Ulu Kelang	5/1/2021	170	Terraced House	Residential	2 - 2 1/2 Storey Terraced	2	NO.30,JALAN WANGSA BUDI 11	WANGSA MELAWATI	1	980000	2021
33765	Kuala Lumpur	Mukim Ulu Kelang	31/5/2017	48	Flat	Residential	Flat	5	23-3-2,JLN AU 3/8	NaN	1	100000	2017

Figure 2.1.1 : Original dataset from NAPIC.

Contents of the Dataset

- **Data Source** : National Property Information Centre (NAPIC) website
- **Data Size** : This dataset comprises records of properties with information on date of valuation, lot size, property type, sector, number of floors, address, price and year.
- **Data Format** : The dataset is available in tabular format with rows representing individual properties and columns representing various attributes.

b. Preliminary Insights

Our preliminary examination of the primary dataset revealed a number of crucial aspects:

- **Data Completeness** : The dataset appears to be relatively complete, with minimal missing values.
- **Data Quality** : Although the dataset appears to be of adequate quality, additional cleaning and preprocessing may be necessary to handle outliers or inconsistencies.
- **Data Features** : The primary dataset includes attributes such as property price, size, and scheme, which will be essential to our analysis.

c. Describing the Dataset

Variable	Type	Anticipated Role	Comments
State	Discrete	Descriptor	State in which the property is located.
District	Discrete	Descriptor	District within the data state where the property is situated.
DateOfValuation	Continuous	Descriptor	Date of property valuation.
LotSize	Continuous	Descriptor	Size of the lot on which the property is built (square metres).
PropertyType	Discrete	Descriptor	Type of property.
Address	Text	Descriptor	Address of property.
Scheme	Discrete	Descriptor	The scheme or development to which the property belongs.
Price	Continuous	Descriptor	The price of the property.
Year	Discrete	Descriptor	The year of data collection.
PricePerMeter	Continuous	Descriptor	It is computed by dividing the property's price by its lot size.

2.2 Supplementary Dataset

d. Dataset Description

The supplementary Dataset is an additional dataset which comes from the same source as primary dataset, but it has additional data which is Rental price and Size. It shows rental price for every scheme and also the size in square feet (sqft) based on Propertyguru and Iproperty.

	Scheme	State	District	PropertyType	Price per sqft	Price per 1000sqft
0	10 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	4.36	4360.0
1	11 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.56	3560.0
2	28 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.55	3550.0
3	ALMASPURI	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.43	2430.0
4	ANGGUNPURI CONDO	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.24	2240.0
...
76	THE NORTHSHORE GARDENS (DESA PARKCITY)	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.34	3340.0
77	TIFFANI KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.57	2570.0
78	VERDANA @ NORTH KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.62	2620.0
79	VILLA ANGSA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	1.60	1600.0
80	VISTA MUTIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	1.25	1250.0

81 rows x 6 columns

Figure 2.2.1 : Supplemental dataset

e. Preliminary Insights

- Data Completeness:** The dataset contains information on various properties in Kuala Lumpur, including their names, locations, property types, and rental prices. Some rows have missing values for the "Rental" and "Size" columns, which may need further investigation or data imputation.
- Data Quality:** Inconsistencies in property names are present (e.g., "10 MONT KIARA" and "11 MONT KIARA"). These might need standardisation. Several properties have missing rental prices and sizes, which could impact the analysis. Property sizes are provided in square feet, but some entries have missing size information. Some properties have identical rental prices, which may be accurate but could also indicate potential errors. Property types seem consistent and are categorised as "Condominium/Apartment."
- Data Features:** Key features include "PropertyType," "Rental," and "Size," which are essential for property analysis. The "State" and "District" columns provide location information within Kuala Lumpur. There is a mix of fully furnished, partially furnished, and unfurnished properties, but this information is not explicitly included in the dataset. The dataset could benefit from additional features such as the number of bedrooms, number of bathrooms, and other amenities.

- **Data Integrity:** There don't appear to be any obvious data integrity issues or duplicates in the dataset, but further data cleaning and validation may be necessary.
- **Data Exploration:** To gain more meaningful insights, further analysis could include examining rental price distributions, property type trends, and location-specific rental patterns. Visualisation techniques can be applied to better understand the data and identify outliers or trends.

f. Describing the Dataset

Variable	Type	Anticipated Role	Comments
Scheme	Discrete	Descriptor	The scheme or development to which the property belongs.
State	Discrete	Descriptor	State in which the property is located.
District	Discrete	Descriptor	District within the data state where the property is situated.
PropertyType	Discrete	Descriptor	Type of property.
Price per sqft	Continuous	Descriptor	The price per sqft for the scheme
Price per 1000sqft	Continuous	Descriptor	The price per sqft times by 1000 for the scheme

3. Data Pre-Processing

In the data pre-processing phase of our analysis, several crucial steps were undertaken to prepare the dataset for further analysis and refine the dataset for subsequent data science tasks. Below is an overview of the key actions performed:

1. Data Loading and Initial Exploration

```
df = pd.read_csv("Data_Science.csv")
df
```

	STATE	DISTRICT	DateOfValuation	LotSize	PropertyType	SECTOR	PRO_TYPE	NoOffFloors	ADDRESS	SCHEME	SYER1	PRICE	YEAR
0	Kuala Lumpur	Kuala Lumpur Town Centre	27/9/2017	65	Condominium/Apartment	Residential	Condominium/Apartment	21	B-8-13,JALAN PUDU ULU	163 SERVICE SUITES	1	375000	2017
1	Kuala Lumpur	Kuala Lumpur Town Centre	28/6/2017	103	Condominium/Apartment	Residential	Condominium/Apartment	21	D-7-3A,JALAN PUDU ULU	163 SERVICE SUITES	1	503000	2017
2	Kuala Lumpur	Kuala Lumpur Town Centre	30/9/2019	104	Condominium/Apartment	Residential	Condominium/Apartment	12	2-10-3,JLN STONOR	1A STONOR APT (CONLAY COURT)	1	660000	2019
3	Kuala Lumpur	Kuala Lumpur Town Centre	1/1/2018	79	Condominium/Apartment	Residential	Condominium/Apartment	8	1-3-5,JLN STONOR	1A STONOR APT (CONLAY COURT)	1	680000	2018
4	Kuala Lumpur	Kuala Lumpur Town Centre	12/8/2018	88	Condominium/Apartment	Residential	Condominium/Apartment	12	1-1-4,JLN STONOR	1A STONOR APT (CONLAY COURT)	1	980000	2018
...
33761	Kuala Lumpur	Mukim Ulu Kelang	25/1/2017	139	Terraced House	Residential	2 - 2 1/2 Storey Terraced	2	75,JLN WANGSA BUDI 7	WANGSA MELAWATI	1	750000	2017
33762	Kuala Lumpur	Mukim Ulu Kelang	7/4/2017	108	Terraced House	Residential	2 - 2 1/2 Storey Terraced	2	4,JALAN 1/24H	WANGSA MELAWATI	1	750000	2017
33763	Kuala Lumpur	Mukim Ulu Kelang	27/8/2019	112	Terraced House	Residential	2 - 2 1/2 Storey Terraced	2	NO. 4,LORONG WANGSA SIAGA 6	WANGSA MELAWATI	1	750000	2019
33764	Kuala Lumpur	Mukim Ulu Kelang	5/1/2021	170	Terraced House	Residential	2 - 2 1/2 Storey Terraced	2	NO.30,JALAN WANGSA BUDI 11	WANGSA MELAWATI	1	980000	2021
33765	Kuala Lumpur	Mukim Ulu Kelang	31/5/2017	48	Flat	Residential	Flat	5	23-3-2,JLN AU 3/8	NaN	1	100000	2017

33766 rows x 13 columns

Figure 3.1.1 : Dataset was sourced from a CSV file, resulting in a DataFrame with 33,766 rows and 13 columns.

- The dataset was initially read from a CSV file and loaded into a Pandas DataFrame.
- Initial inspection of the data set revealed that it contained 33,766 rows and 13 columns. The columns in the dataset included information such as state, district, date of valuation, lot size, property type, number of floors, address, scheme, price and year.

2. Filtering Out Unnecessary Data

	STATE	DISTRICT	DateOfValuation	LotSize	PRO_TYPE	NoOffFloors	ADDRESS	SCHEME	PRICE	YEAR
4676	Kuala Lumpur	Mukim Batu	28/11/2018	323.1	Condominium/Apartment	22	AA-12-02,JALAN KIARA 1	10 MONT KIARA	2600000	2018
4677	Kuala Lumpur	Mukim Batu	7/6/2019	344.1	Condominium/Apartment	22	B-8-2,JALAN KIARA 1	10 MONT KIARA	2600000	2019
4678	Kuala Lumpur	Mukim Batu	4/6/2020	337.4	Condominium/Apartment	22	AA-15-01,JALAN KIARA 1	10 MONT KIARA	2700000	2020
4679	Kuala Lumpur	Mukim Batu	7/1/2021	323.1	Condominium/Apartment	22	AA-13-02,JALAN KIARA 1	10 MONT KIARA	2700000	2021
4680	Kuala Lumpur	Mukim Batu	20/1/2020	337.4	Condominium/Apartment	22	A-09-01,JALAN KIARA 1	10 MONT KIARA	2800000	2020
...
12520	Kuala Lumpur	Mukim Batu	27/3/2017	118.076	Condominium/Apartment	23	1-13A-07,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000	2017
12521	Kuala Lumpur	Mukim Batu	27/3/2017	118.076	Condominium/Apartment	23	1-13A-07,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000	2017
12522	Kuala Lumpur	Mukim Batu	15/10/2020	173	Condominium/Apartment	23	1-23-05,JLN 1/12D, OFF BATU 4 1/2, JLN IPOH	ZETA DESKYE ALAM SAUJANA	800000	2020
12523	Kuala Lumpur	Mukim Batu	13/4/2018	23391	Vacant Plot	NaN	26685,JLN JINJANG UTARA	NaN	35522665	2018
12524	Kuala Lumpur	Mukim Batu	13/4/2018	23788	Vacant Plot	NaN	26684,JLN JINJANG UTARA	NaN	36125569	2018

7849 rows x 10 columns

Figure 3.2.1 : Filtering out unnecessary data based on the DISTRICT.

- Additionally, data records pertaining to the district known as "Mukim Batu" were considered irrelevant for the purposes of the analysis and were subsequently excluded from the dataset.
- The removal of 7,849 rows was achieved by generating a boolean mask using the 'DISTRICT' column and subsequently dropping the rows that satisfied the given condition.

	STATE	DISTRICT	DateOfValuation	LotSize	PRO_TYPE	NoOfFloors	ADDRESS	SCHEME	PRICE	YEAR
4676	Kuala Lumpur	Mukim Batu	28/11/2018	323.1	Condominium/Apartment	22	AA-12-02,JALAN KIARA 1	10 MONT KIARA	2600000	2018
4677	Kuala Lumpur	Mukim Batu	7/6/2019	344.1	Condominium/Apartment	22	B-8-2,JALAN KIARA 1	10 MONT KIARA	2600000	2019
4678	Kuala Lumpur	Mukim Batu	4/6/2020	337.4	Condominium/Apartment	22	AA-15-01,JALAN KIARA 1	10 MONT KIARA	2700000	2020
4679	Kuala Lumpur	Mukim Batu	7/1/2021	323.1	Condominium/Apartment	22	AA-13-02,JALAN KIARA 1	10 MONT KIARA	2700000	2021
4680	Kuala Lumpur	Mukim Batu	20/1/2020	337.4	Condominium/Apartment	22	A-09-01,JALAN KIARA 1	10 MONT KIARA	2800000	2020
...
12518	Kuala Lumpur	Mukim Batu	16/5/2019	116	Condominium/Apartment	23	1-17-03A,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	600000	2019
12519	Kuala Lumpur	Mukim Batu	2/11/2017	118.076	Condominium/Apartment	23	N-19-7,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	698000	2017
12520	Kuala Lumpur	Mukim Batu	27/3/2017	118.076	Condominium/Apartment	23	1-13A-07,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000	2017
12521	Kuala Lumpur	Mukim Batu	27/3/2017	118.076	Condominium/Apartment	23	1-13A-07,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000	2017
12522	Kuala Lumpur	Mukim Batu	15/10/2020	173	Condominium/Apartment	23	1-23-05,JLN 1/12D, OFF BATU 4 1/2, JLN IPOH	ZETA DESKYE ALAM SAUJANA	800000	2020

4280 rows x 10 columns

Figure 3.2.2 : Filtering out unnecessary data based on the PRO_TYPE.

- Another refinement was implemented to exclude records associated with the property type 'Condominium/Apartment'.
- Using a similar method, a boolean mask was created based on the 'PRO_TYPE' column, and rows meeting the condition were removed. This operation resulted in the removal of 4,280 rows.

3. Removing Unnecessary Columns

	LotSize	PropertyType	NoOfFloors	Address	Scheme	Price	Year
1	323.1	Condominium/Apartment	22	AA-12-02,JALAN KIARA 1	10 MONT KIARA	2600000	2018
2	344.1	Condominium/Apartment	22	B-8-2,JALAN KIARA 1	10 MONT KIARA	2600000	2019
3	337.4	Condominium/Apartment	22	AA-15-01,JALAN KIARA 1	10 MONT KIARA	2700000	2020
4	323.1	Condominium/Apartment	22	AA-13-02,JALAN KIARA 1	10 MONT KIARA	2700000	2021
5	337.4	Condominium/Apartment	22	A-09-01,JALAN KIARA 1	10 MONT KIARA	2800000	2020

Figure 3.3.1: After removing unnecessary columns.

- In order to improve the cleanliness and relevance of the dataset, we have identified certain columns, namely 'PropertyType', 'SECTOR', 'SYER1', 'STATE', 'DISTRICT', 'NoOfFloors' and 'DateOfValuation', as being unnecessary for our analysis.
- These columns have been removed from the DataFrame, resulting in a 7-columned dataset.

4. Column Renaming

LotSize	PropertyType	Address	Scheme	Price	Year
---------	--------------	---------	--------	-------	------

Figure 3.4.1: Renaming column to improve the readability of the data set.

- The DataFrame column names were renamed to improve clarity and consistency. In particular, the columns 'PRO_TYPE' to 'PropertyType,' 'ADDRESS' to 'Address,' 'SCHEME' to 'Scheme,' 'PRICE' to 'Price,' and 'YEAR' to 'Year.' This action was taken to make the column names more descriptive and easier to work with.

5. Checking for Missing Values

```
LotSize      0
PropertyType  0
NoOfFloors   1
Address      0
Scheme       0
Price        0
Year         0
dtype: int64
```

Figure 3.5.1: NoOfFloors column had one missing value.

- a. Using the "isnull()" function, it was determined whether the dataset contained missing values (NaN). As demonstrated, the sum of missing values in each column is zero, further confirming that there are no missing values in any of the columns within the dataset.

6. Data Type Conversion

- a. Certain columns' data types were modified to correspond with their respective data. The 'LotSize' and 'Price' columns were converted to floating-point numbers using the "astype(float)" method, guaranteeing that these values are treated as decimals. 'Year' are simultaneously converted to 64-bit integers (int64) using astype(np.int64). This transformation ensured that the dataset treated 'NoOfFloors' and 'Year' as integers.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4280 entries, 1 to 4280
Data columns (total 6 columns):
#   Column      Non-Null Count  Dtype
---  -
0   LotSize     4280 non-null  float64
1   PropertyType 4280 non-null  object
2   Address     4280 non-null  object
3   Scheme      4280 non-null  object
4   Price       4280 non-null  float64
5   Year        4280 non-null  int64
dtypes: float64(2), int64(1), object(3)
memory usage: 200.8+ KB
```

Figure 3.6.1: Data Type Conversion

7. Handling Infinite Values

```
# Count how many division errors or infinite values there are in PricePerSize
np.isinf(df_b['PricePerSize']).sum()

17
```

Figure 3.7.1: Result how many infinite values in the PricePerMeter column.

	LotSize	PropertyType	Address	Scheme	Price	Year	PricePerMeter
77	0.0	Condominium/Apartment	D-18-11,JALAN KIARA 1	11 MONT KIARA	2200000.0	2017	inf
167	0.0	Condominium/Apartment	A1-32-2,JALAN KIARA	28 MONT KIARA	1940000.0	2017	inf
664	0.0	Condominium/Apartment	A-1-1,JLN DUTAMAS ORKID	CONCERTO NORTH KIARA	1270000.0	2017	inf
842	0.0	Condominium/Apartment	B-06-07,JALAN DUTA HARTAMAS	HARTAMAS REGENCY 2	860000.0	2017	inf
1001	0.0	Condominium/Apartment	A-7-3A,JALAN KIARA 3	KIARA 3	650000.0	2017	inf
1007	0.0	Condominium/Apartment	A-17-2,JALAN KIARA 3	KIARA 3	670000.0	2018	inf
1299	0.0	Condominium/Apartment	B-19-03,JALAN DESA KIARA	KIARAMAS SUTERA KONDOMINIUM	1100000.0	2017	inf
2160	0.0	Condominium/Apartment	A-2-2,NO.1,PERSIARAN RESIDEN,PHASE 1B,	NADIA PARKFRONT CONDO (DESA PARKCITY)	880000.0	2017	inf
2302	0.0	Condominium/Apartment	30-3,CHANGKAT DUTA KIARA	ONE CENTRAL PARK KONDO	1725000.0	2017	inf
2550	0.0	Condominium/Apartment	C-2-4,JLN 1/18B,TMN BATU PERMAI,	PERMAI RIA CONDOMINIUM	180000.0	2018	inf
2662	0.0	Condominium/Apartment	A-6-10,JALAN PRIMA PELANGI 1	PRIMA PELANGI	390000.0	2017	inf
2689	0.0	Condominium/Apartment	16-9,JALAN 6/38A	PRIMA TIARA	360000.0	2019	inf
2857	0.0	Condominium/Apartment	B-36-03,JALAN KIARA 4	RESIDENSI 22 MONT KIARA	2000000.0	2018	inf
3559	0.0	Condominium/Apartment	A-27-3A,JALAN KUCHING	SRI PUTRAMAS III (ROYAL REGENT)	1090000.0	2017	inf
3683	0.0	Condominium/Apartment	22-2,PERSIARAN RESIDEN	THE WESTSIDE ONE (DESA PARKCITY)	1080000.0	2017	inf
3923	0.0	Condominium/Apartment	A-36-03,CHANGKAT DUTA KIARA	TIFFANI KIARA	4000000.0	2017	inf
3951	0.0	Condominium/Apartment	A1-18-5,JLN DUTAMAS MELATI	VERDANA @ NORTH KIARA	890000.0	2017	inf

Figure 3.7.2: Which rows have infinite values.

- In 'PricePerMeter,' we have found 17 infinite values that suggest division errors or incorrect values.

	LotSize	PropertyType	Address	Scheme	Price	Year	PricePerMeter
1	323.100	Condominium/Apartment	AA-12-02,JALAN KIARA 1	10 MONT KIARA	2600000.0	2018	8047.044259
2	344.100	Condominium/Apartment	B-8-2,JALAN KIARA 1	10 MONT KIARA	2600000.0	2019	7555.943040
3	337.400	Condominium/Apartment	AA-15-01,JALAN KIARA 1	10 MONT KIARA	2700000.0	2020	8002.371073
4	323.100	Condominium/Apartment	AA-13-02,JALAN KIARA 1	10 MONT KIARA	2700000.0	2021	8356.545961
5	337.400	Condominium/Apartment	A-09-01,JALAN KIARA 1	10 MONT KIARA	2800000.0	2020	8298.755187
...
4276	116.000	Condominium/Apartment	1-17-03A,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	600000.0	2019	5172.413793
4277	118.076	Condominium/Apartment	N-19-7,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	698000.0	2017	5911.446865
4278	118.076	Condominium/Apartment	1-13A-07,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000.0	2017	6182.458755
4279	118.076	Condominium/Apartment	1-13A-07,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000.0	2017	6182.458755
4280	173.000	Condominium/Apartment	1-23-05,JLN 1/12D, OFF BATU 4 1/2, JLN IPOH	ZETA DESKYE ALAM SAUJANA	800000.0	2020	4624.277457

4263 rows x 7 columns

Figure 3.7.3: Dropping Inf Values

- In this case, we opted to remove rows containing inf values to maintain data consistency and facilitate subsequent analysis.

8. Duplicate Row Removal

```
duplicate_rows_df = df_partA[df_partA.duplicated()]
print("number of duplicate rows: ", duplicate_rows_df.shape)

number of duplicate rows: (68, 7)
```

Figure 3.8.1: Result number of duplicate rows.

- We have identified and counted the total number of duplicate rows in the DataFrame, which totals 68. To fix this, we utilised the "drop_duplicates" method to eliminate these duplicate rows.

	LotSize	PropertyType	Address	Scheme	Price	Year	PricePerMeter
1	323.100	Condominium/Apartment	AA-12-02,JALAN KIARA 1	10 MONT KIARA	2600000.0	2018	8047.044259
2	344.100	Condominium/Apartment	B-8-2,JALAN KIARA 1	10 MONT KIARA	2600000.0	2019	7555.943040
3	337.400	Condominium/Apartment	AA-15-01,JALAN KIARA 1	10 MONT KIARA	2700000.0	2020	8002.371073
4	323.100	Condominium/Apartment	AA-13-02,JALAN KIARA 1	10 MONT KIARA	2700000.0	2021	8356.545961
5	337.400	Condominium/Apartment	A-09-01,JALAN KIARA 1	10 MONT KIARA	2800000.0	2020	8298.755187
...
4275	89.741	Condominium/Apartment	2-6-7,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	580000.0	2017	6463.043648
4276	116.000	Condominium/Apartment	1-17-03A,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	600000.0	2019	5172.413793
4277	118.076	Condominium/Apartment	N-19-7,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	698000.0	2017	5911.446865
4278	118.076	Condominium/Apartment	1-13A-07,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000.0	2017	6182.458755
4280	173.000	Condominium/Apartment	1-23-05,JLN 1/12D, OFF BATU 4 1/2, JLN IPOH	ZETA DESKYE ALAM SAUJANA	800000.0	2020	4624.277457

4195 rows x 7 columns

Figure 3.8.2: Result after dropping duplicate rows.

- The resulting dataset contained 4,195 rows, each of which was unique and distinct. Eliminating duplicate rows is a crucial step for ensuring data consistency and preventing redundancy, as it enables accurate and trustworthy analysis and modelling with a more concise dataset.

9. Detect and Remove Outliers

	LotSize	PropertyType	Address	Scheme	Price	Year	PricePerMeter
10	29730.000	Condominium/Apartment	AA-28-02,JALAN KIARA 1	10 MONT KIARA	2880000.0	2018	96.871847
32	379.900	Condominium/Apartment	BA-07-01,JALAN KIARA 1	10 MONT KIARA	3400000.0	2019	8949.723611
33	379.900	Condominium/Apartment	B-12-1,JALAN KIARA 1	10 MONT KIARA	3450000.0	2020	9081.337194
34	344.100	Condominium/Apartment	BA-40-02,JALAN KIARA 1	10 MONT KIARA	3450000.0	2018	10026.155187
35	379.900	Condominium/Apartment	BB-12-01,JALAN KIARA 1	10 MONT KIARA	3488000.0	2017	9181.363517
...
3183	68.282	Condominium/Apartment	C-11-01,CHANGKAT DUTA KIARA	SENI MONT KIARA	5500000.0	2017	80548.314343
3184	735.000	Condominium/Apartment	D-11-3A,CHANGKAT DUTA KIARA	SENI MONT KIARA	5600000.0	2019	7619.047619
3185	827.000	Condominium/Apartment	A-40-3,CHANGKAT DUTA KIARA	SENI MONT KIARA	7833760.0	2018	9472.503023
3835	41.000	Condominium/Apartment	07-01,JALAN RESIDEN UTAMA	THE WESTSIDE TWO(DESAPARKCITY)	1468500.0	2017	35817.073171
4148	23.000	Condominium/Apartment	A-16-05,JLN PRIMA PELANGI 7	VILLA ORKID	640000.0	2018	27826.086957

61 rows x 7 columns

Figure 3.9.1 : Outlier Detection.

- We employed the Z-score method, a commonly used statistical technique, to detect outliers. We successfully identified a total of 61 rows exhibiting outlier values within the dataset.

	LotSize	PropertyType	Address	Scheme	Price	Year	PricePerMeter
1	323.100	Condominium/Apartment	AA-12-02,JALAN KIARA 1	10 MONT KIARA	2600000.0	2018	8047.044259
2	344.100	Condominium/Apartment	B-8-2,JALAN KIARA 1	10 MONT KIARA	2600000.0	2019	7555.943040
3	337.400	Condominium/Apartment	AA-15-01,JALAN KIARA 1	10 MONT KIARA	2700000.0	2020	8002.371073
4	323.100	Condominium/Apartment	AA-13-02,JALAN KIARA 1	10 MONT KIARA	2700000.0	2021	8356.545961
5	337.400	Condominium/Apartment	A-09-01,JALAN KIARA 1	10 MONT KIARA	2800000.0	2020	8298.755187
...
4275	89.741	Condominium/Apartment	2-6-7,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	580000.0	2017	6463.043648
4276	116.000	Condominium/Apartment	1-17-03A,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	600000.0	2019	5172.413793
4277	118.076	Condominium/Apartment	N-19-7,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	698000.0	2017	5911.446865
4278	118.076	Condominium/Apartment	1-13A-07,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000.0	2017	6182.458755
4280	173.000	Condominium/Apartment	1-23-05,JLN 1/12D, OFF BATU 4 1/2, JLN IPOH	ZETA DESKYE ALAM SAUJANA	800000.0	2020	4624.277457

4134 rows x 7 columns

Figure 3.9.1 : After removing outliers.

- To enhance the integrity of the dataset, we removed the rows identified as outliers from the original DataFrame. This cleaned dataset includes 4,134 rows and retains the original seven columns.

10. Changing Row Number Index

	LotSize	PropertyType	Address	Scheme	Price	Year	PricePerMeter
1	323.100	Condominium/Apartment	AA-12-02,JALAN KIARA 1	10 MONT KIARA	2600000.0	2018	8047.044259
2	344.100	Condominium/Apartment	B-8-2,JALAN KIARA 1	10 MONT KIARA	2600000.0	2019	7555.943040
3	337.400	Condominium/Apartment	AA-15-01,JALAN KIARA 1	10 MONT KIARA	2700000.0	2020	8002.371073
4	323.100	Condominium/Apartment	AA-13-02,JALAN KIARA 1	10 MONT KIARA	2700000.0	2021	8356.545961
5	337.400	Condominium/Apartment	A-09-01,JALAN KIARA 1	10 MONT KIARA	2800000.0	2020	8298.755187
...
4130	89.741	Condominium/Apartment	2-6-7,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	580000.0	2017	6463.043648
4131	116.000	Condominium/Apartment	1-17-03A,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	600000.0	2019	5172.413793
4132	118.076	Condominium/Apartment	N-19-7,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	698000.0	2017	5911.446865
4133	118.076	Condominium/Apartment	1-13A-07,JLN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000.0	2017	6182.458755
4134	173.000	Condominium/Apartment	1-23-05,JLN 1/12D, OFF BATU 4 1/2, JLN IPOH	ZETA DESKYE ALAM SAUJANA	800000.0	2020	4624.277457

4134 rows x 7 columns

Figure 3.10.1 : Changing Row Number Index starts from 1 to 4134.

- To improve the dataset's readability, the row index numbers were adjusted beginning with 1.
- The total number of rows in the DataFrame was determined using the "len()" function. Using "pd.RangeIndex", the row index was reset to begin at 1 and extend to the total number of rows.

4. Data Pre-Processing for Supplementary Dataset

1. Data Loading and Data Presentation

Defines the location (path) where a CSV file is stored. This path is where the code will look for the CSV file. It takes the path from path and combines it with the name of the CSV file ("new_data.csv"). This creates the full file path, which specifies the exact location of the file on the computer. It uses the pandas library, a popular data manipulation tool in Python, to read the data from the CSV file specified by the full file path. The data is loaded into a structure called a DataFrame, which is like a table of data. After reading the data, it displays the contents of the DataFrame. This allows us to see and work with the data in a structured format. can perform various operations on this data, such as analysing, processing, or visualising it.

Unnamed: 0		State	District	DateOfValuation	LotSize	PropertyType	NoOffFloors	Address	Scheme	Price	Year	PricePerSize
0	1	Kuala Lumpur	Mukim Batu	28/11/2018	323.100	Condominium/Apartment	22	AA-12-02,JALAN KIARA 1	10 MONT KIARA	2600000	2018	8047.044259
1	2	Kuala Lumpur	Mukim Batu	7/6/2019	344.100	Condominium/Apartment	22	B-8-2,JALAN KIARA 1	10 MONT KIARA	2600000	2019	7555.943040
2	3	Kuala Lumpur	Mukim Batu	4/6/2020	337.400	Condominium/Apartment	22	AA-15-01,JALAN KIARA 1	10 MONT KIARA	2700000	2020	8002.371073
3	4	Kuala Lumpur	Mukim Batu	7/1/2021	323.100	Condominium/Apartment	22	AA-13-02,JALAN KIARA 1	10 MONT KIARA	2700000	2021	8356.545961
4	5	Kuala Lumpur	Mukim Batu	20/1/2020	337.400	Condominium/Apartment	22	A-09-01,JALAN KIARA 1	10 MONT KIARA	2800000	2020	8298.755187
...
4275	4276	Kuala Lumpur	Mukim Batu	16/5/2019	116.000	Condominium/Apartment	23	1-17-03A,JILN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	600000	2019	5172.413793
4276	4277	Kuala Lumpur	Mukim Batu	2/11/2017	118.076	Condominium/Apartment	23	N-19-7,JILN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	698000	2017	5911.446865
4277	4278	Kuala Lumpur	Mukim Batu	27/3/2017	118.076	Condominium/Apartment	23	1-13A-07,JILN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000	2017	6182.458755
4278	4279	Kuala Lumpur	Mukim Batu	27/3/2017	118.076	Condominium/Apartment	23	1-13A-07,JILN KG BT 5, OFF JLN IPOH	ZETA DESKYE ALAM SAUJANA	730000	2017	6182.458755
4279	4280	Kuala Lumpur	Mukim Batu	15/10/2020	173.000	Condominium/Apartment	23	1-23-05,JLN 1/12D, OFF BATU 4 1/2, JLN IPOH	ZETA DESKYE ALAM SAUJANA	800000	2020	4624.277457
4280 rows x 12 columns												

Figure 4.1.1: Data Loading and Data Presentation

2. Selecting and Dropping Columns

It identifies a list of column names (columns_to_drop) that you want to remove or exclude from the DataFrame. These columns are not needed for the analysis or are redundant. The code removes the columns listed in columns_to_drop from the original DataFrame (df). This operation creates a new DataFrame called df_pb, which stands for "DataFrame after dropping specified columns." The displays the first few rows of the cleaned DataFrame (df_pb). This allows you to see the DataFrame without the columns you chose to drop.

	State	District	DateOfValuation	PropertyType	Scheme
0	Kuala Lumpur	Mukim Batu	28/11/2018	Condominium/Apartment	10 MON'T KIARA
1	Kuala Lumpur	Mukim Batu	7/6/2019	Condominium/Apartment	10 MON'T KIARA
2	Kuala Lumpur	Mukim Batu	4/6/2020	Condominium/Apartment	10 MON'T KIARA
3	Kuala Lumpur	Mukim Batu	7/1/2021	Condominium/Apartment	10 MON'T KIARA
4	Kuala Lumpur	Mukim Batu	20/1/2020	Condominium/Apartment	10 MON'T KIARA
...
4275	Kuala Lumpur	Mukim Batu	16/5/2019	Condominium/Apartment	ZETA DESKYE ALAM SAUJANA
4276	Kuala Lumpur	Mukim Batu	2/11/2017	Condominium/Apartment	ZETA DESKYE ALAM SAUJANA
4277	Kuala Lumpur	Mukim Batu	27/3/2017	Condominium/Apartment	ZETA DESKYE ALAM SAUJANA
4278	Kuala Lumpur	Mukim Batu	27/3/2017	Condominium/Apartment	ZETA DESKYE ALAM SAUJANA
4279	Kuala Lumpur	Mukim Batu	15/10/2020	Condominium/Apartment	ZETA DESKYE ALAM SAUJANA

4280 rows x 5 columns

Figure 4.2.1: Selecting and Dropping Columns

3. Grouping Data by 'Scheme' Column and Selecting the First Row of Each Group

groups the data in the DataFrame df based on the values in the 'Scheme' column. This operation is done using the group by method. It groups the rows with the same 'Scheme' value together. For each group of rows with the same 'Scheme' value, the code selects the first row. This is done using the first() method. So, after this step, you have a DataFrame where each 'Scheme' appears only once, and the data for each 'Scheme' corresponds to the first row with that 'Scheme' value. It reduces the rows by 4280 rows to 135 rows.

	Scheme	State	District	DateOfValuation	PropertyType
0	10 MON'T KIARA	Kuala Lumpur	Mukim Batu	28/11/2018	Condominium/Apartment
1	11 MONT KIARA	Kuala Lumpur	Mukim Batu	21/8/2020	Condominium/Apartment
2	28 MONT KIARA	Kuala Lumpur	Mukim Batu	27/11/2020	Condominium/Apartment
3	ALAM PURI 51	Kuala Lumpur	Mukim Batu	26/4/2021	Condominium/Apartment
4	ALMASPURI	Kuala Lumpur	Mukim Batu	29/11/2019	Condominium/Apartment
...
130	VILLA MAKMUR CONDOMINIUM	Kuala Lumpur	Mukim Batu	3/2/2017	Condominium/Apartment
131	VILLA ORKID	Kuala Lumpur	Mukim Batu	16/1/2019	Condominium/Apartment
132	VISTA KIARA	Kuala Lumpur	Mukim Batu	26/10/2020	Condominium/Apartment
133	VISTA MUTIARA	Kuala Lumpur	Mukim Batu	22/5/2017	Condominium/Apartment
134	ZETA DESKYE ALAM SAUJANA	Kuala Lumpur	Mukim Batu	5/8/2019	Condominium/Apartment

135 rows x 5 columns

Figure 4.3.1: Grouping Data by 'Scheme' Column and Selecting the First Row of Each Group

4. Exporting the DataFrame to a CSV File

Exports the DataFrame `unique_schemes_df` to a CSV file named 'new_data.csv'. This operation is done using the `to_csv` method provided by pandas. The DataFrame is saved to a CSV file with the specified name.

```
unique_schemes_df.to_csv('new_data.csv')

from google.colab import files
files.download('supplementary_dataset.csv')
```

Figure 4.4.1: Exporting the DataFrame to a CSV File

5. Adding New Column Data Manually from internet source.

Manually adding new data of rental price for each scheme from Iproperty and Propertyguru.

PropertyGuru

Buy Rent Condos New Launches Commercial Find Agent Guides More


Antah tower

Filter 3


Apartment / Condo /

53867 Other Properties For Rent


View All



RM 1,600 /mo
3 2
Arte Cheras
Jalan Midah 2, Cheras, Kuala Lumpur
Service Residence



RM 1,700 /mo
3 2
Arte Cheras
Jalan Midah 2, Cheras, Kuala Lumpur
Service Residence



RM 2,300 /mo
2 2
Trion @ KL
Lot 162 Jalan Sungai Besi, Pudu, Kuala Lumpur
Service Residence

iProperty

Buy Rent New Launches Transactions Advertise Guides & Insights Events

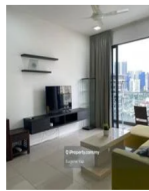
Home > Rent

All States

Anjali north kiara

Residential property for rent in Malaysia with Anjali north kiara

All Residential Min Rent (RM) Max Rent (RM) 750 sq. ft. - 1500 sq. ft. Bedrooms



Eugene Yap Posted on 21 Sep 2023 04:32 PM
RM 3,100 (RM 2.35 per sq. ft.)
Anjali North Kiara, Segambut
Segambut, Kuala Lumpur
Condominium • Built-up : 1,319 sq. ft. • Fully furnished
3 2 2
Save View details

1	Scheme	State	District	PropertyType	Price per sqft
2	10 MON'T KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartn	4.36
3	11 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartn	3.56
4	28 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartn	3.55
5	ALAM PURI 51	Kuala Lumpur	Mukim Batu	Condominium/Apartment	
6	ALMASPURI	Kuala Lumpur	Mukim Batu	Condominium/Apartn	2.43
7	AMANDARI KONDO	Kuala Lumpur	Mukim Batu	Condominium/Apartment	
8	ANGGUNPURI CONDO	Kuala Lumpur	Mukim Batu	Condominium/Apartn	2.24
9	ANGKUPURI	Kuala Lumpur	Mukim Batu	Condominium/Apartn	2.43
10	ANGSA APT	Kuala Lumpur	Mukim Batu	Condominium/Apartment	

Figure 4.5.1: Adding New Column Data Manually from internet source.

6. Loading New CSV file and Display

Loads new data that has already been added from a CSV file into a DataFrame. The file is specified by the variable `csv_url`, which should contain the correct path or URL to the CSV file.

	Scheme	State	District	PropertyType	Price per sqft
0	10 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	4.36
1	11 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.56
2	28 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.55
3	ALAM PURI 51	Kuala Lumpur	Mukim Batu	Condominium/Apartment	NaN
4	ALMASPURI	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.43
...
130	VILLA MAKMUR CONDOMINIUM	Kuala Lumpur	Mukim Batu	Condominium/Apartment	NaN
131	VILLA ORKID	Kuala Lumpur	Mukim Batu	Condominium/Apartment	NaN
132	VISTA KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	NaN
133	VISTA MUTIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	1.25
134	ZETA DESKYE ALAM SAUJANA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	NaN
135 rows x 5 columns					

Figure 4.6.1: Loading New CSV file and Display.

7. Dropping Rows with Missing Values

Drops rows in the DataFrame `df` where there are missing values (NaN) in the 'Price per sqft' columns. This is done using the `dropna` method with the `subset` parameter specifying the columns to consider ('Price per sqft'). The `inplace=True` argument means that the changes are made directly to the DataFrame `df` without creating a new DataFrame. It reduces the rows from 135 rows to 81 rows.

	Scheme	State	District	PropertyType	Price per sqft
0	10 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	4.36
1	11 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.56
2	28 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.55
3	ALMASPURI	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.43
4	ANGGUNPURI CONDO	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.24
...
76	THE NORTHSHORE GARDENS (DESA PARKCITY)	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.34
77	TIFFANI KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.57
78	VERDANA @ NORTH KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.62
79	VILLAANGSANA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	1.60
80	VISTA MUTIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	1.25
81 rows x 5 columns					

Figure 4.7.1: Dropping Rows with Missing Values.

8. Calculate 'Price per 1000sqft'

The 'Price per 1000sqft' column in the DataFrame is created by taking the 'Price per 1000sqft' values and multiplying them by 1000, effectively expressing the property price in terms of a 1000 square foot area. This can be useful for comparing property prices across different properties or for standardising the price metric to a consistent unit of area.

	Scheme	State	District	PropertyType	Price per sqft	Price per 1000sqft
0	10 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	4.36	4360.0
1	11 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.56	3560.0
2	28 MONT KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.55	3550.0
3	ALMASPURI	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.43	2430.0
4	ANGGUNPURI CONDO	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.24	2240.0
...
76	THE NORTHSHORE GARDENS (DESA PARKCITY)	Kuala Lumpur	Mukim Batu	Condominium/Apartment	3.34	3340.0
77	TIFFANI KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.57	2570.0
78	VERDANA @ NORTH KIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	2.62	2620.0
79	VILLA ANGSA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	1.60	1600.0
80	VISTA MUTIARA	Kuala Lumpur	Mukim Batu	Condominium/Apartment	1.25	1250.0
81 rows x 6 columns						

Figure 4.7.1: Dropping Rows with Missing Values.

5. Exploratory Data Analysis

1. Summary Statistics

In the first phase of our Exploratory Data Analysis (EDA), we conducted a thorough examination of the dataset, focusing on summary statistics for the variables of interest: LotSize, Price, and PricePerMeter. These summary statistics provide crucial insights into the data's central tendencies, dispersion, and overall distribution.

	LotSize	Price	PricePerMeter
count	4134.000000	4.134000e+03	4134.000000
mean	153.950354	1.015199e+06	6231.664434
std	76.674293	6.682745e+05	2352.646974
min	36.000000	2.500000e+04	246.434099
25%	102.000000	4.700000e+05	4308.811559
50%	132.970000	8.400000e+05	6053.706024
75%	186.000000	1.430000e+06	8044.672920
max	1460.000000	3.350000e+06	15092.375748

Figure 5.1.1: Summary statistics for the EDA results.

- **Count:** The number of non-missing (non-null) values in each numerical column.
- **Mean:** The average value of each numerical column.
- **Std:** The standard deviation, which measures the spread or dispersion of data.
- **Min:** The minimum value in each numerical column.
- **25%:** The 25th percentile value (also known as the first quartile).
- **50%:** The median (50th percentile) value, which represents the middle value when data is sorted.
- **75%:** The 75th percentile value (also known as the third quartile).
- **Max:** The maximum value in each numerical column

- **LotSize**

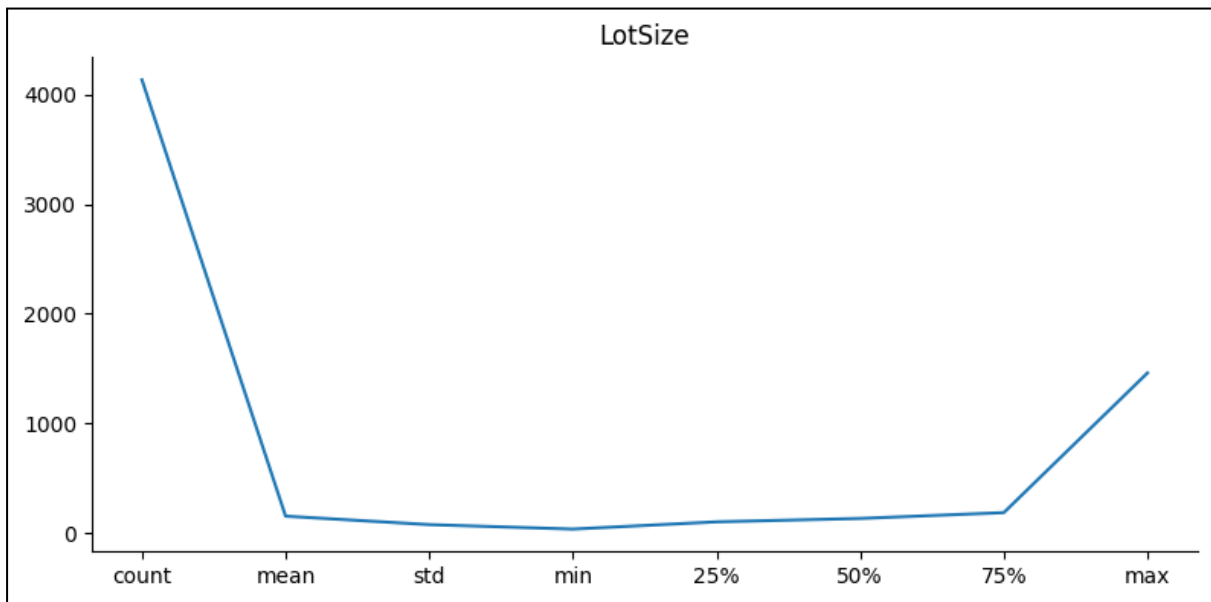


Figure 5.1.2: Line Plot Values Graph for LotSize.

The LotSize variable of the dataset provides insightful information about the size distribution of properties. The average lot size is approximately 153.95 metres, with a standard deviation of approximately 76.67 metres, indicating some variation in lot sizes. There is a wide range of lot sizes, from a minimum of 36 metres to a maximum of 1,460 metres. The percentiles provide additional granularity: 25% of lot sizes are less than or equal to 102 metres, and 75% are less than or equal to 186 metres. The median lot size, which represents the middle value, measures around 132.97 metres.

- **Price**

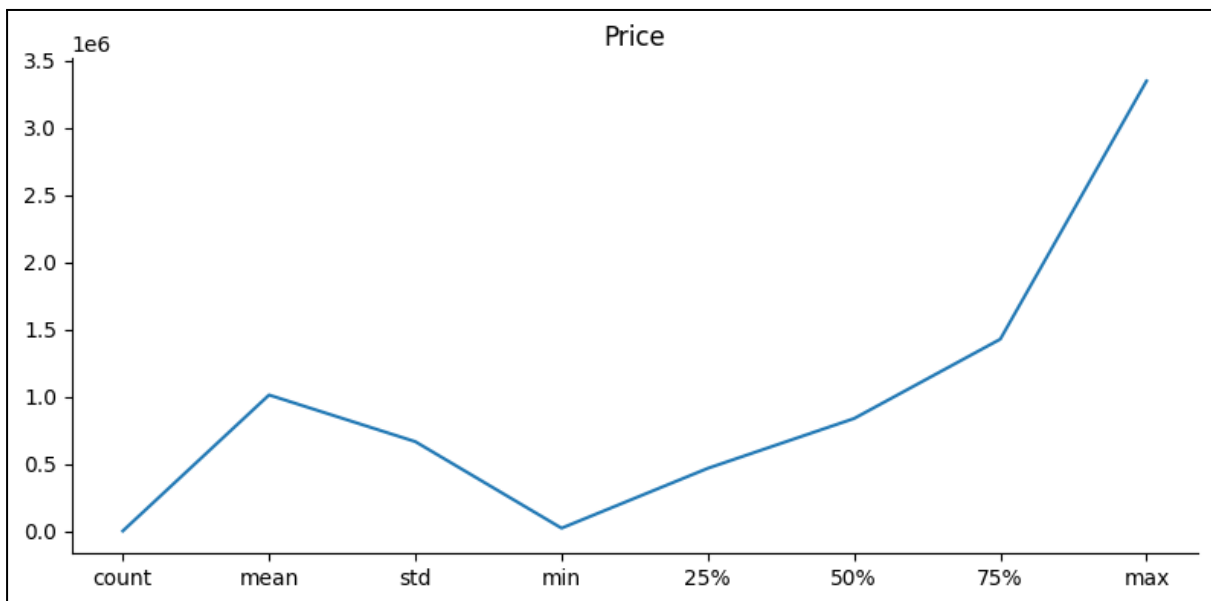


Figure 5.1.3: Line Plot Values Graph for Price

The dataset contains 4,134 property price observations, providing a comprehensive view of the pricing landscape. The average property price is around RM 1,015,199, with a standard deviation of about RM 668,274.50, indicating a significant degree of price variability. The price range for real estate is wide, ranging from a low of RM 25,000 to a high of RM 3,350,000. Examining the percentiles reveals more about the price distribution: 25% of the properties are priced at or below RM 470,000, while 75% are priced at RM 1,430,000 or less. The median property price, which represents the middle value, is around RM 840,000.

- **PricePerMeter**

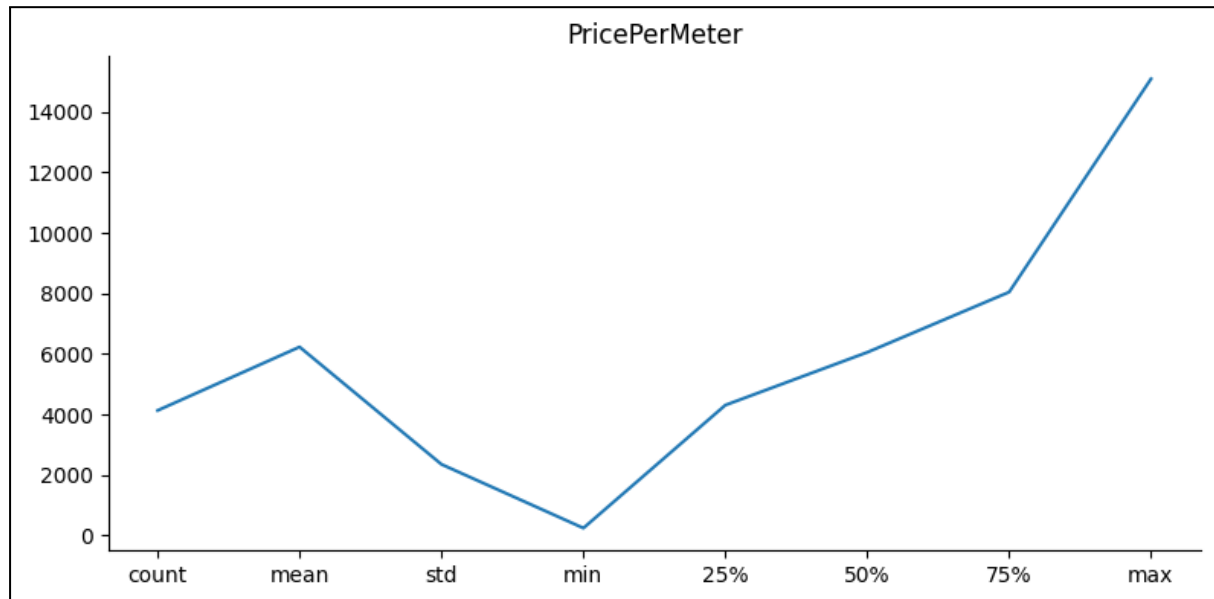


Figure 5.1.4: Line Plot Values Graph for PricePerMeter.

The dataset contains 4,134 price per square metre observations, offering information on the cost patterns associated with property size. The average price per square metre is about RM 6,231.66, with a standard deviation of around RM 2,352.65, demonstrating substantial variation in this parameter. The price per square metre range is quite wide, with a minimum of about RM 246.43 and a maximum of about RM 15,092.38.

Percentiles provide additional insight into the price per square metre distribution: 25% of the values are at or below RM 4,308.81, with the remaining 75% falling within a range of RM 8,044.67 or less. The median price per square metre, or the middle value, is around RM 6,053.71.

2. Question

We will start the Exploratory Data Analysis for each stated question by using the cleaned dataset as follow;

- PartB.csv
 - 4134 rows, 7 columns (LotSize, PropertyType, NoOfFloors, Address, Scheme, Price, Year, PricePerMeter).

2.1 Descriptive Question

2.1.1 What is the trend of property price for condominiums in the Mukim Batu for the last 5 years of 2017 until 2021? Plot a graph for the trend for overall condominium in the Mukim Batu.

To answer the question, “What is the trend of property price for condominiums in the Mukim Batu for the last 5 years of 2017 until 2021? Plot a graph for the trend for overall condominium in the Mukim Batu.”, we can conduct an Exploratory Data Analysis (EDA) focusing on the condominium properties located in Mukim Baru. Here’s how we can approach this:

First, the dataset is grouped by 'Year' in order to analyse the annual trend in condo property prices in Mukim Batu. By grouping the data annually, we are able to observe the changes and fluctuations in property prices over time and gain insights regarding annual trends. Next, the mean 'Price' and 'PricePerMeter' are calculated for each year to provide a comprehensive overview of the average property prices and prices per metre over the specified time period. By averaging the price and price per metre, it is possible to observe general trends and shifts over time in Mukim Batu's real estate market without the distraction of extreme values.

	Year	Price
0	2017	9.881340e+05
1	2018	1.037752e+06
2	2019	1.021554e+06
3	2020	1.013909e+06
4	2021	1.001809e+06

Figure 5.2.1.1.1: Average price condominium properties from 2017-2021.

	Year	PricePerMeter
0	2017	6083.287612
1	2018	6318.534202
2	2019	6316.867969
3	2020	6189.843908
4	2021	6283.509645

Figure 5.2.1.1.2: Average price per metre condominium properties from 2017-2021.

➤ Data Visualization

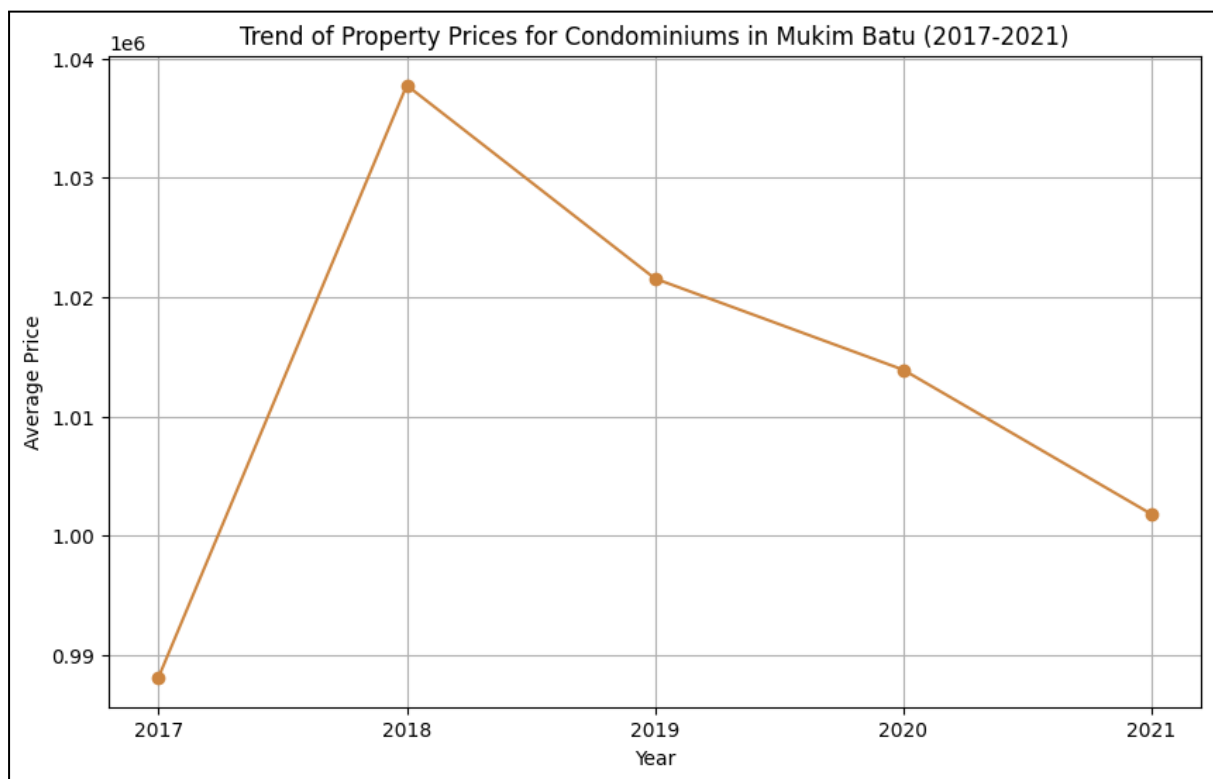


Figure 5.2.1.1.3: Line Graph for trend of condominium prices in Mukim Batu (2017-2021).

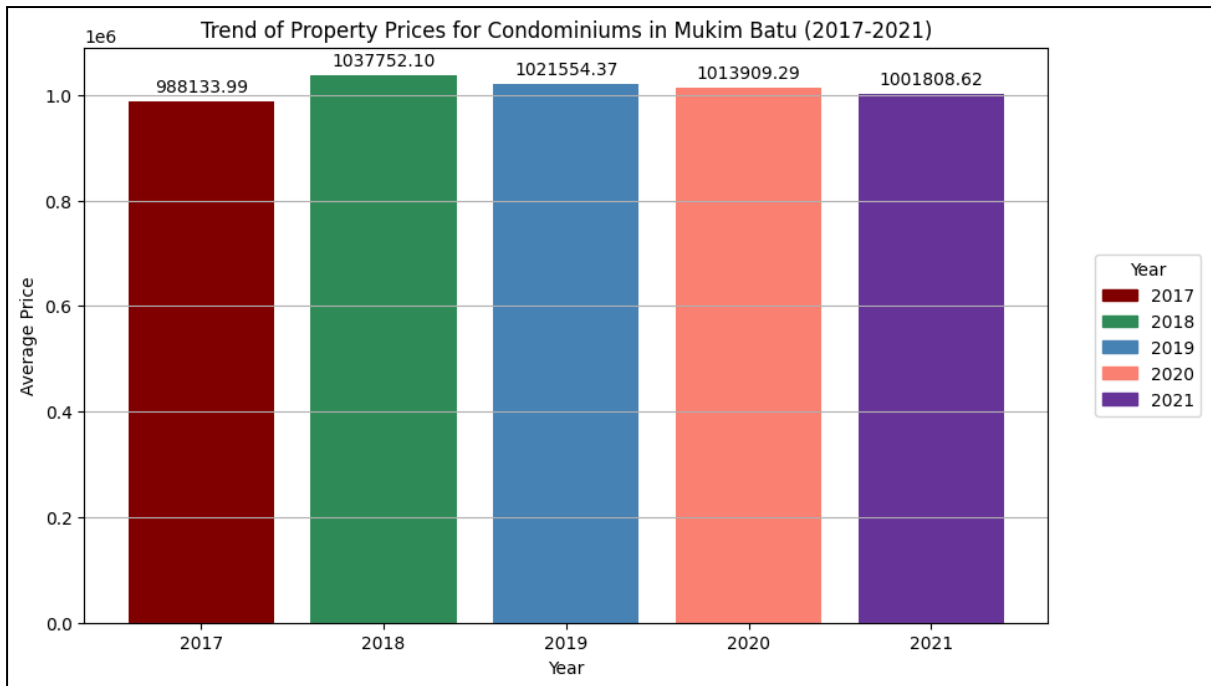


Figure 5.2.1.1.4: Bar Graph for trend of condominium prices in Mukim Batu (2017-2021).

The year with the highest average price is 2018 with 1037752.10
The year with the lowest average price is 2017 with 988133.99

Figure 5.2.1.1.5: Interpretation of trend of condominium prices in Mukim Batu (2017-2021).

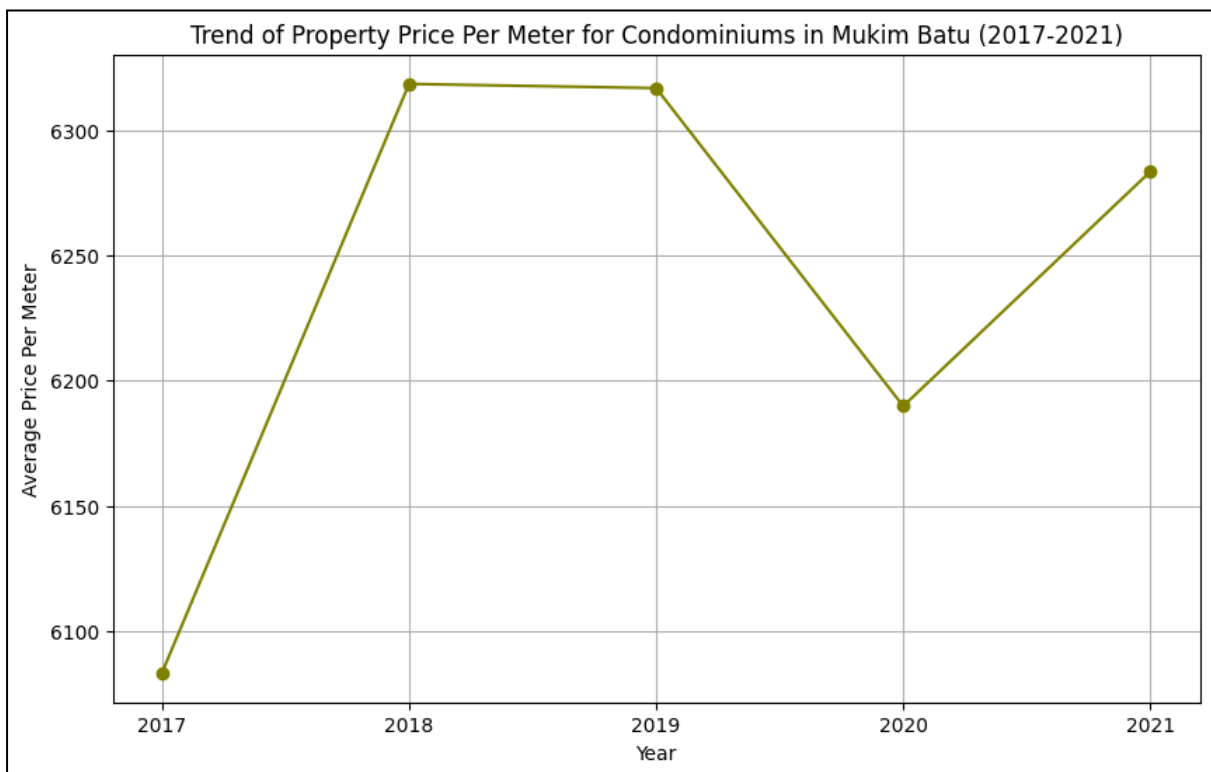


Figure 5.2.1.1.6: Line Graph for trend of condominium price per metre in Mukim Batu (2017-2021)

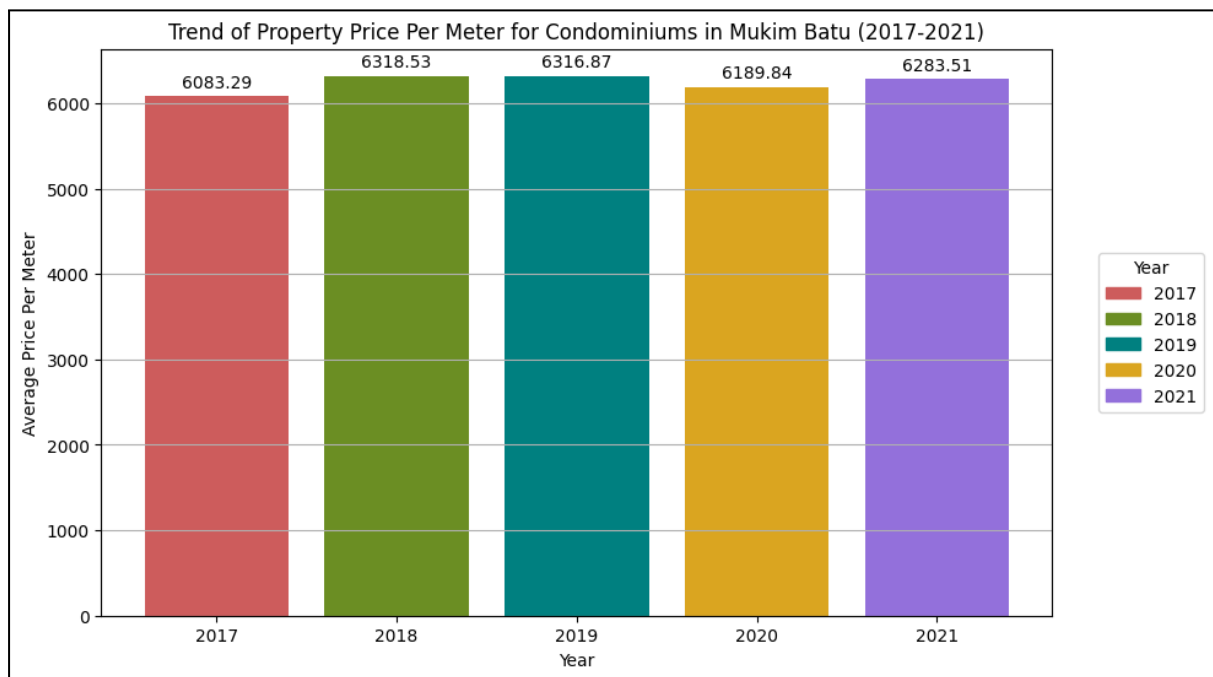


Figure 5.2.1.1.7: Bar Graph for trend of median condominium price per metre in Mukim Batu (2017-2021).

The year with the highest average price per meter is 2018 with 6318.53
 The year with the lowest average price per meter is 2017 with 6083.29

Figure 5.3.1.1.8: Interpretation of trend of condominium price per metre in Mukim Batu (2017-2021).

In 2018, the average property price increased from approximately 988,134 MYR in 2017 to approximately 1,037,752 MYR. However, after 2018, there was a gradual decline in average prices, culminating in an average price of approximately 1,001,809 MYR in 2021.

The average price per metre fluctuated, peaking in 2018 at approximately 6318.53 MYR/m. Following 2018, there was a slight decrease, with an average price per metre of approximately 6283.51 MYR/m in 2021.

Property prices appeared to be stabilising, with minor fluctuations, according to the trends. The initial rise in average prices and price per metre was followed by a gradual decline, indicating a balanced market in Mukim Batu.

2.1.2 What is the median property price for condominiums in the Mukim Batu from 2017 until 2021?

To answer the question, "What is the median property price for condominiums in the Mukim Batu from 2017 until 2021?", we can conduct an exploratory data analysis (EDA) focusing on condominium properties located in Mukim Batu. Here's how we can approach this:

The data was grouped by year to determine the median property price and median price per metre for condominiums in Mukim Batu from 2017 to 2021, and the median value for 'Price' and 'PricePerMeter' was calculated for each grouped year. This method aids in providing a clear and concise view of the trends in median property prices and price per metre over the specified time period, and the median, as a measure of central tendency, is robust to outliers and extreme values in the dataset, providing a reliable insight into market conditions.

	Year	Price
0	2017	782500.0
1	2018	875000.0
2	2019	880000.0
3	2020	800000.0
4	2021	830000.0

Figure 5.2.1.2.1: Median price condominium properties from 2017-2021.

	Year	PricePerMeter
0	2017	6031.883770
1	2018	6309.148265
2	2019	6101.877833
3	2020	5628.887291
4	2021	5982.905983

Figure 5.2.1.2.2: Median price per metre condominium properties from 2017-2021

The preceding data represents the median condo prices and price per size in Mukim Batu between 2017 and 2021.

➤ **Data Visualization**

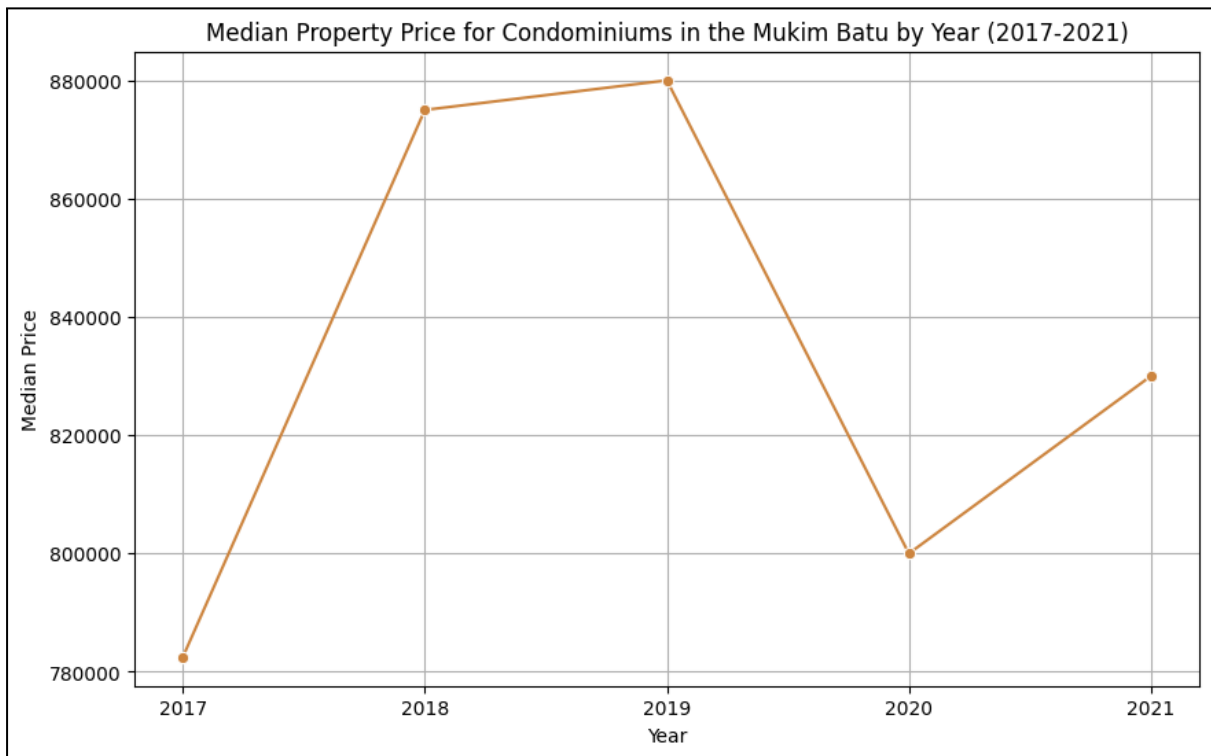


Figure 5.2.1.2.3: Line Graph for median of condominium prices in Mukim Batu (2017-2021).

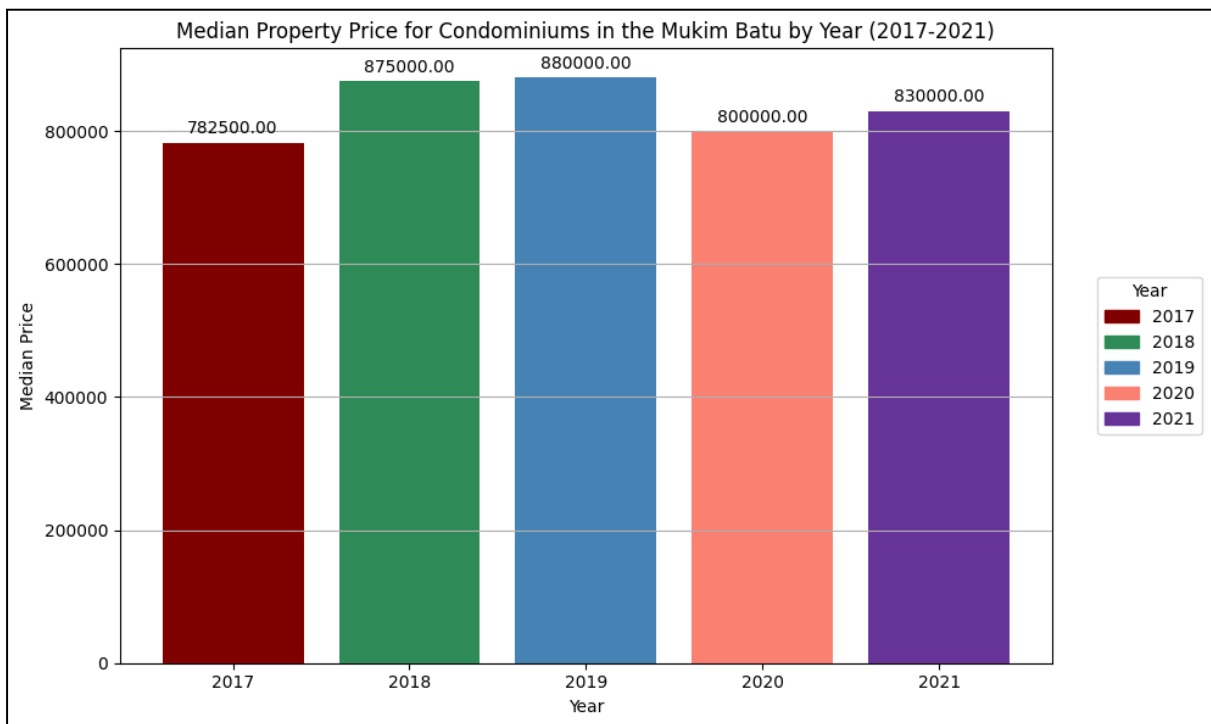


Figure 5.2.1.2.4: Bar Graph for median of condominium prices in Mukim Batu (2017-2021).

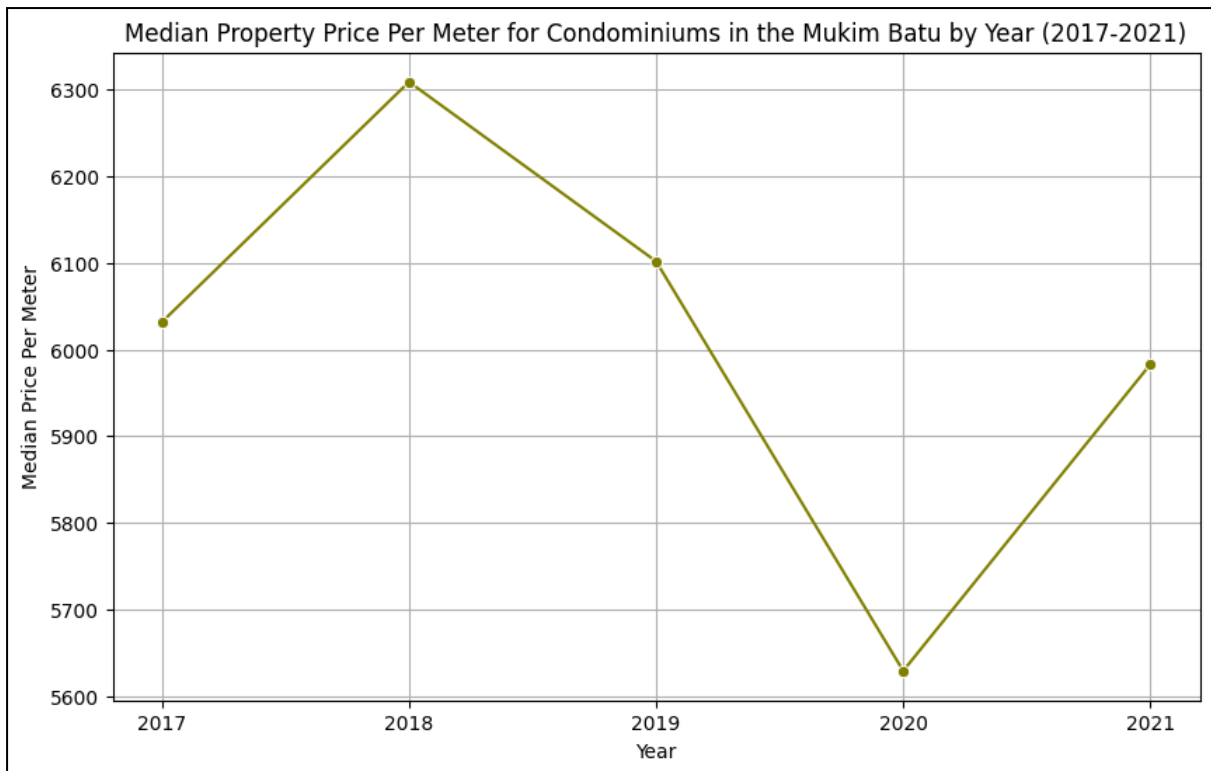


Figure 5.2.1.2.5: Line Graph for median of condominium price per metre in Mukim Batu (2017-2021).

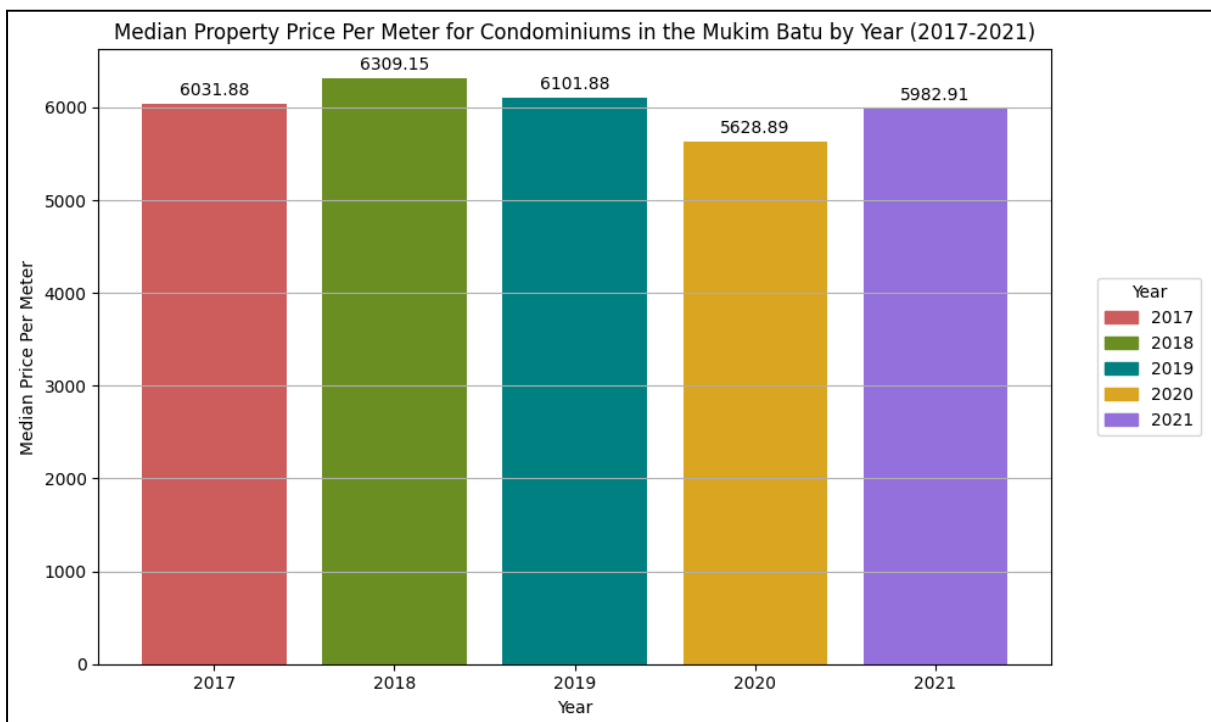


Figure 5.2.1.2.6: Bar Graph for median of condominium price per meter in Mukim Batu (2017-2021).

From 2017 to 2021, the Mukim Batu condominium market experienced several fluctuations in median property prices and median prices per metre. In 2017, the median property price was 782,500 MYR, with a price per metre of around 6031.88 MYR/m. In the following year, 2018, both metrics increased, with the median property price reaching 875,000 MYR and the median price per metre rising to around 6309.15 MYR/m. However, while the median property price increased slightly in 2019 to 880,000 MYR, the median price per metre decreased marginally to approximately 6101.88 MYR/m. The median property price and median price per metre fell significantly in 2020, falling to 800,000 MYR and approximately 5628.89 MYR/m, respectively. Nonetheless, the median property price increased to 830,000 MYR in 2021, with the median price per metre returning to around 5982.91 MYR/m. These variations over time reflect the dynamic nature of the Mukim Batu real estate market, which may be influenced by the interplay of supply and demand, economic conditions, and policy changes in the region.

3.1.3 What is the median Price Per Metre for each condominium scheme?

To gain insight into the median price per unit size (PricePerMeter) for each condominium scheme in our dataset, we will first group the data by the "Scheme" variable. The "Scheme" variable represents distinct condominium projects or schemes within our dataset. By grouping the data in this manner, we can isolate each scheme and calculate the median PricePerMeter for each scheme.

We will then determine the median PricePerMeter for each scheme. The median is a reliable statistic that represents the middle value of an ascendingly sorted dataset. In this context, it will provide a measure of central tendency for PricePerMeter within each condominium scheme, allowing us to determine the typical pricing per unit size for condominiums within a particular scheme.

By analysing and comparing the median PricePerMeter values of various condominium schemes, we are able to determine the variation in pricing structures and trends. This analysis can provide buyers and real estate professionals with valuable insights, allowing them to determine which schemes may offer favourable price-to-size ratios and which may have a higher median Price Per Meter.

	Scheme	MedianPricePerMeter
0	10 MON'T KIARA	8887.312633
1	11 MONT KIARA	8703.094598
2	28 MONT KIARA	7883.330296
3	ALAM PURI 51	5528.846154
4	ALMASPURI	6597.222222

Figure 5.2.1.3.1: DataFrame for MedianPricePerMeter by each scheme.

For a comprehensive view of the entire table, including median Price Per Metre values for all schemes in the dataset, please refer to the following link: [Full Table - Median Price Per Meter](#).

To gain insight into the pricing dynamics of various condominium schemes, we determined the median price per unit size (PricePerMeter) for each scheme in the dataset. The partial results presented above display the median Price Per Meter for select schemes.

➤ Data Visualization

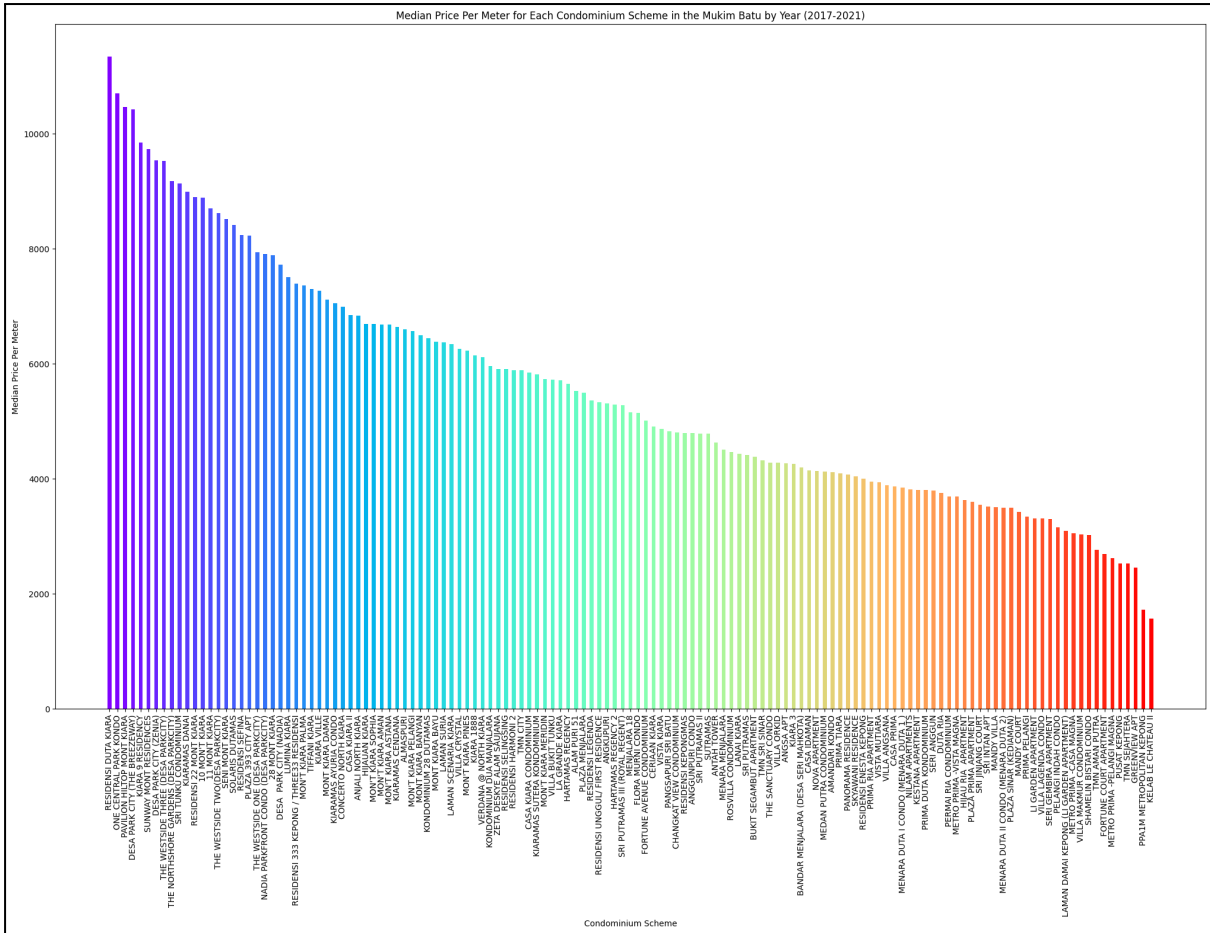


Figure 5.2.1.3.2: Bar Plot Median Price Per Meter for Each Condominium Scheme

For a more detailed and visually clear representation of these schemes, please refer to the full images via this link: [Full Visualizations - Median Price Per Meter](#).

The 10 highest schemes by median price per size are:

	Scheme	MedianPricePerMeter
92	RESIDENSI DUTA KIARA	11337.426877
74	ONE CENTRAL PARK KONDO	10704.000000
77	PAVILION HILTOP MONT KIARA	10464.890099
21	DESA PARK CITY (THE BREEZEWAY)	10416.666667
36	KIARA 9 RESIDENCY	9841.980000
113	SUNWAY MONT RESIDENCES	9725.490196
22	DESA PARK CITY (ZENIA)	9529.702970
118	THE WESTSIDE THREE (DESA PARKCITY)	9529.380718
115	THE NORTHSHORE GARDENS (DESA PARKCITY)	9176.470588
112	SRI TUNKU CONDOMINIUM	9137.055838

The 10 lowest schemes by median price per size are:

	Scheme	MedianPricePerMeter
32	KELAB LE CHATEAU II	1564.102564
84	PPA1M METROPOLITAN KEPONG	1726.618705
27	GREENVIEW APT	2454.991817
123	TMN SEJAHTERA	2522.524631
89	PUSAT KEPONG	2524.030403
59	METRO PRIMA -PELANGI MAGNA	2617.647059
26	FORTUNE COURT APARTMENT	2688.235294
121	TMN AMAN PUTRA	2761.194030

Figure 5.2.1.3.3: The 10 highest and 10 lowest schemes by median Price Per Meter.

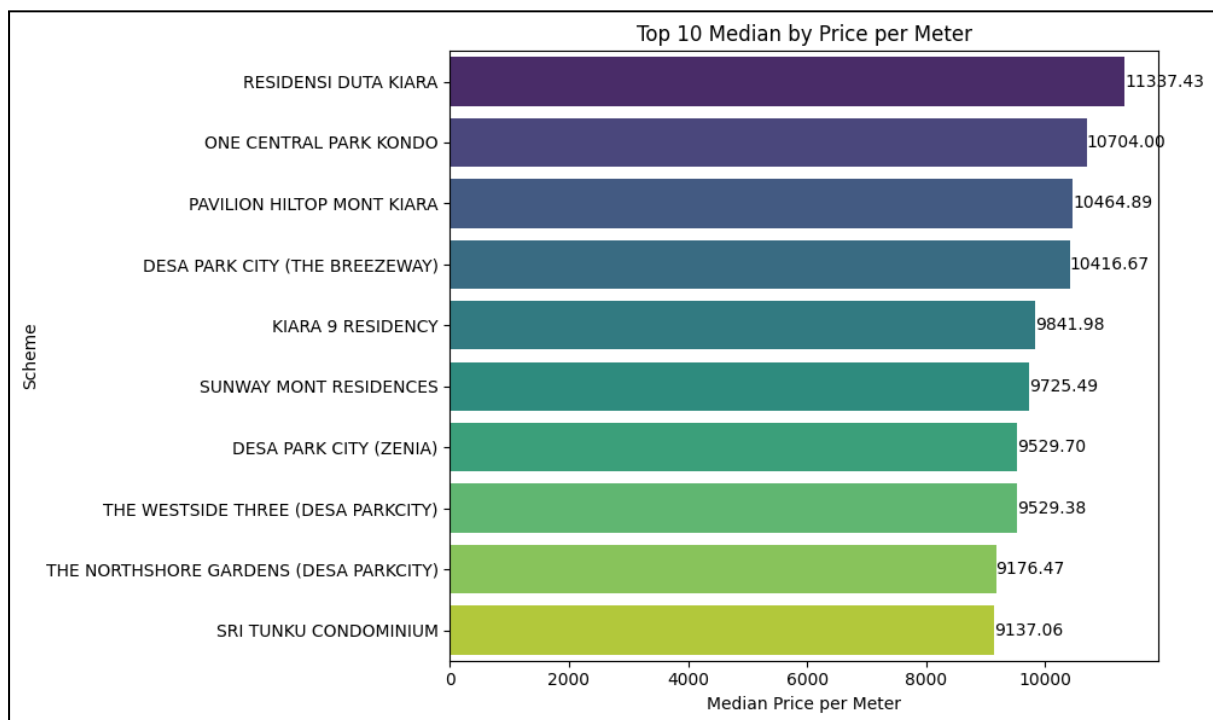


Figure 5.2.1.3.4: The 10 highest by median Price Per Meter.

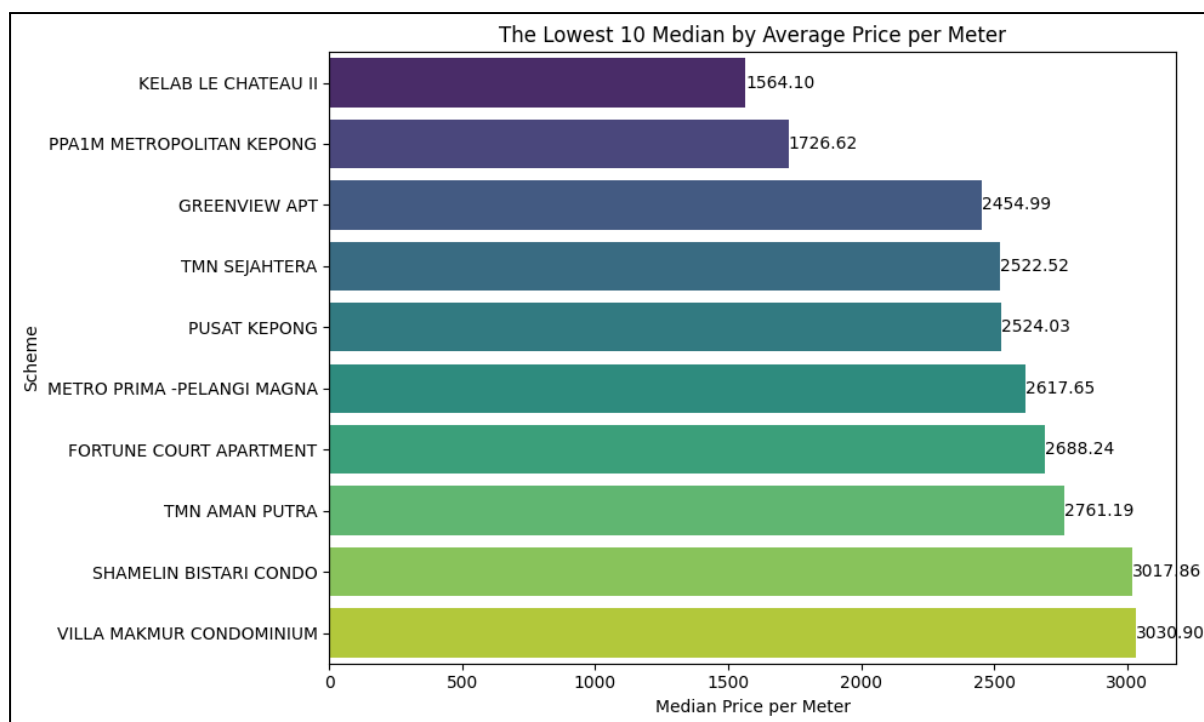


Figure 5.2.1.3.4: The 10 lowest by median Price Per Meter.

The picture above is a bar graph comparing the median prices of various condominium schemes. The x-axis indicates the name of the condominium scheme, whereas the y-axis indicates the median Price Per Meter. The bars are coloured in a rainbow gradient, with purple representing the median price that is highest and red representing the median price that is lowest. The bars are arranged in descending order of median price from left to right.

Our objective in analysing the data visualisation was to identify the different pricing dynamics among the different condominium schemes. In order to do this, we computed each scheme's median price per unit of size (PricePerMeter), which offers important insights into the relative pricing structures of various projects. With a median Price Per Meter of roughly RM 11,337.43, "RESIDENSI DUTA KIARA" stands out as the scheme with the highest price among the findings. This suggests a premium pricing structure and that the size and amenities of the condominiums included in this scheme are valued. Plans like "ONE CENTRAL PARK KONDO" and "PAVILION HILTOP MONT KIARA," which exhibit premium and competitive pricing, are closely vying with each other. Conversely, programmes like "KELAB LE CHATEAU II" and "PPA1M METROPOLITAN KEPONG" emphasise affordability by providing the lowest median prices per size. For prospective purchasers and real estate agents looking to comprehend affordability and pricing patterns within particular condominium complexes, these insights are priceless.

3.1.4 What is the average Price Per Meter for each condominium scheme?

To gain insight into the pricing dynamics of condominiums within various schemes, we calculate the average Price Per Meter using the mean function. Using the mean function, we determine the average price for each condominium scheme in the dataset. This method allows us to determine the average price per unit area for each scheme, providing buyers and sellers with valuable information about the real estate market.

By calculating the average Price Per Meter, we can determine pricing trends across various schemes. Such information enables prospective purchasers to evaluate the affordability of condominiums in various schemes, while sellers can gain a deeper understanding of how their offerings compare to those of their competitors.

	Scheme	AveragePricePerMeter
0	10 MONT KIARA	8825.628485
1	11 MONT KIARA	8773.533354
2	28 MONT KIARA	8018.119120
3	ALAM PURI 51	5421.739202
4	ALMASPURI	6551.930964

Figure 5.2.1.4.1: DataFrame for Average PricePerMeter by each scheme.

For a comprehensive view of the entire table, including average Price Per Meter values for all schemes in the dataset, please refer to the following link: [Full Table - Average Price Per Meter](#).

In order to gain insight into the pricing dynamics of various condominium schemes, we calculated the average price per unit size (PricePerMeter) for each scheme in our dataset. The partial results are displayed above, displaying the average Price Per Meter for select schemes.

➤ Data Visualization

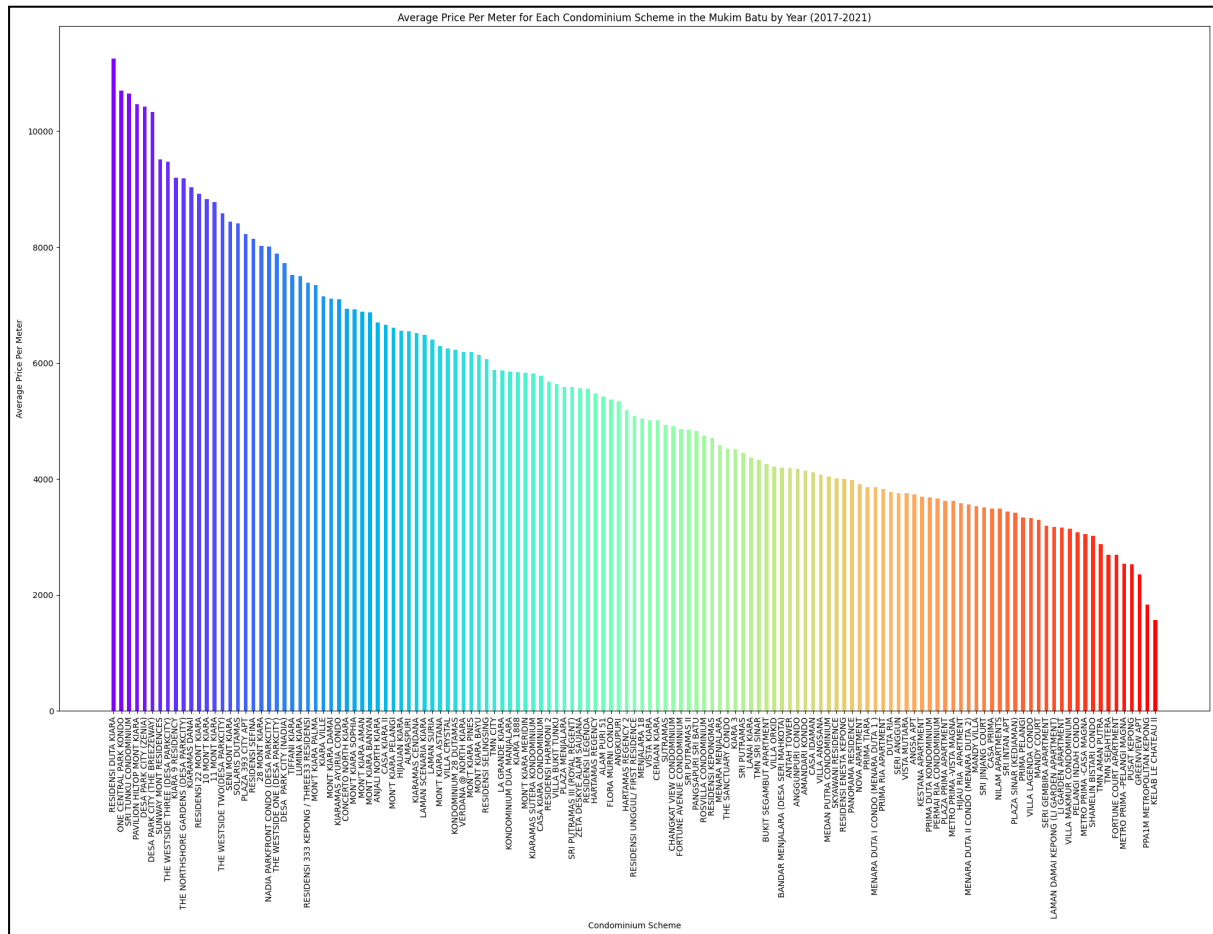


Figure 5.2.1.4.2: Bar Plot Average Price Per Meter for Each Condominium Scheme.

For a more detailed and visually clear representation of these schemes, please refer to the full images via this link: [Full Visualizations - Average Price Per Meter](#).

The 10 highest schemes by average price per meter are:		
	Scheme	AveragePricePerMeter
92	RESIDENSI DUTA KIARA	11252.004408
74	ONE CENTRAL PARK KONDO	10702.681825
112	SRI TUNKU CONDOMINIUM	10654.031989
77	PAVILION HILLTOP MONT KIARA	10461.191891
22	DESA PARK CITY (ZENIA)	10422.645276
21	DESA PARK CITY (THE BREEZEWAY)	10331.857498
113	SUNWAY MONT RESIDENCES	9514.587974
118	THE WESTSIDE THREE (DESA PARKCITY)	9474.680580
36	KIARA 9 RESIDENCY	9198.954702
115	THE NORTHSHORE GARDENS (DESA PARKCITY)	9190.922484
The 10 lowest schemes by average price per meter are:		
	Scheme	AveragePricePerMeter
32	KELAB LE CHATEAU II	1564.102564
84	PPA1M METROPOLITAN KEPONG	1829.670053
27	GREENVIEW APT	2351.247558
89	PUSAT KEPONG	2524.030403
59	METRO PRIMA -PELANGI MAGNA	2540.639496
26	FORTUNE COURT APARTMENT	2688.235294
123	TMN SEJAHTERA	2693.460351
121	TMN AMAN PUTRA	2879.541686
104	SHAMELIN BISTARI CONDO	3017.857143
58	METRO PRIMA -CASA MAGNA	3052.631579

Figure 5.2.1.4.3: The 10 highest and 10 lowest schemes by average Price Per Meter

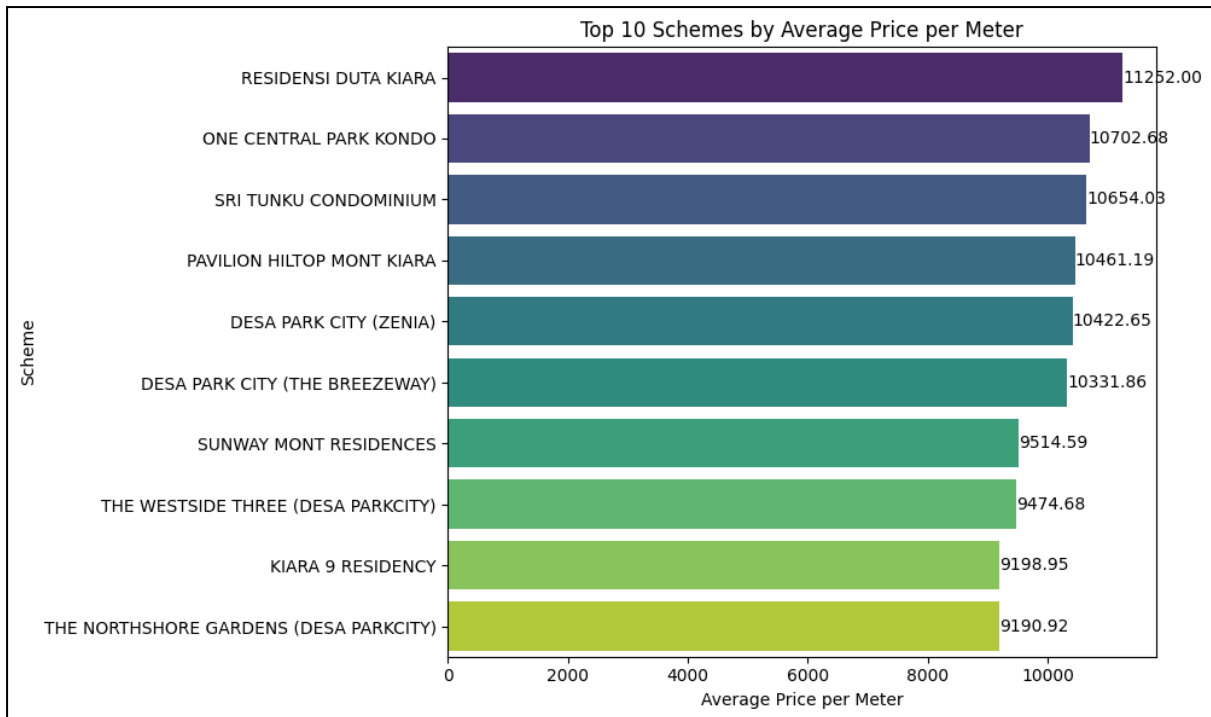


Figure 5.2.1.4.4: The top 10 highest schemes by average Price Per Meter

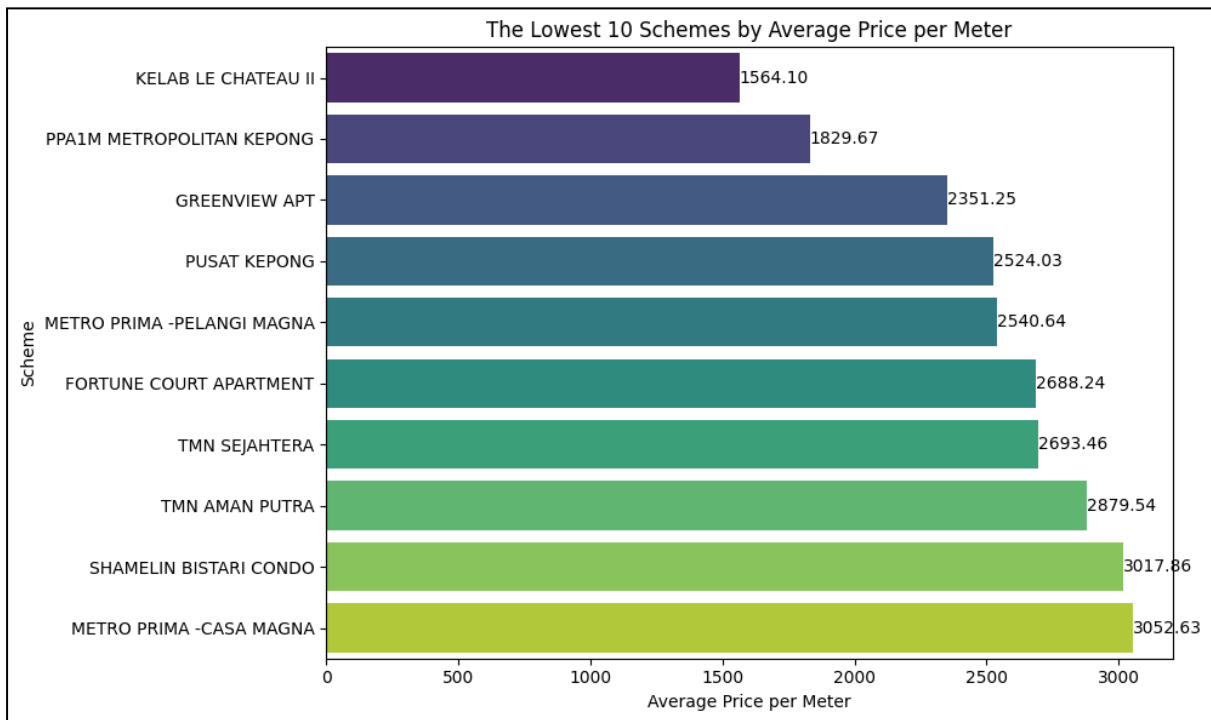


Figure 5.2.1.4.5: The top 10 lowest schemes by average Price Per Meter

The picture above is a bar graph comparing the average prices of various condominium schemes. The x-axis indicates the name of the condominium scheme, whereas the y-axis indicates the average Price Per Meter. The bars are gradiently coloured, with purple representing the highest average price and red representing the lowest. The bars are arranged in descending order of average price, from left to right.

The evaluation of condominium schemes in Mukim Batu revealed that the average price per metre varied significantly between schemes. Top-tier projects such as "Residensi Duta Kiara," "One Central Park Kondo," and "Sri Tunku Condominium" displayed significantly higher average prices per metre, surpassing the 10,000 MYR/m threshold, with "Residensi Duta Kiara" being the most expensive at 11,252.00 MYR/m. These premium prices were reflected in other prestigious projects, such as "Pavilion Hilltop Mont Kiara" and "Desa Park City (Zenia)," indicating a superior market valuation. In contrast, "Kelab Le Chateau II," "PPA1M Metropolitan Kepong," and "Greenview Apt" represented the lower end of the market, with average prices per metre hovering around 1,564.10 MYR/m, 1,829.67 MYR/m, and 2,351.25 MYR/m respectively. Other condominium schemes in the region, such as "Pusat Kepong" and "Metro Prima -Pelangi Magna," also exhibited subdued prices per metre, reflecting a wide range of valuations for condominium schemes. The observed differences in the average prices per metre likely reflect the diverse offerings, locations, and amenities of the various projects, highlighting the multifaceted condominium market in Mukim Batu.

3.2 Causal Question

3.2.1 Does lot size influence the price of a condominium?

Table: Price vs Calculated Price (based on LotSize and PricePerMeter)					
	LotSize	PricePerMeter	Price	CalculatedPrice	PriceApproxEqualsCalculated
0	323.100	8047.044259	2600000.0	2600000.0	True
1	344.100	7555.943040	2600000.0	2600000.0	True
2	337.400	8002.371073	2700000.0	2700000.0	True
3	323.100	8356.545961	2700000.0	2700000.0	True
4	337.400	8298.755187	2800000.0	2800000.0	True
...
4129	89.741	6463.043648	580000.0	580000.0	True
4130	116.000	5172.413793	600000.0	600000.0	True
4131	118.076	5911.446865	698000.0	698000.0	True
4132	118.076	6182.458755	730000.0	730000.0	True
4133	173.000	4624.277457	800000.0	800000.0	True
4134 rows x 5 columns					
Total number of False values in PriceApproxEqualsCalculated: 0					

Figure 3.2.1.1.1: Table of Price vs Calculated Price (Based on LotSize and PricePerMeter) and total number of false values in Price ApproxEqualsCalculated

The provided table presents a comparative analysis between the 'Price' column and a derived price value, obtained by multiplying the 'LotSize' by the 'PricePerMeter' for each entry in a given dataset. The column labelled 'PriceApproxEqualsCalculated' will provide an indication of whether the 'Price' value is approximately equal to the 'CalculatedPrice' value for each individual row, while considering the specified tolerance. The 'PriceApproxEqualsCalculated' variable contains no false values, indicating that all price values are true. Once the examination of the overall count of erroneous values has been completed, the subsequent step involves the use of data visualisation techniques in order to analyse the correlation between LotSize and Price.

	LotSize	Price
count	4134.000000	4.134000e+03
mean	153.950354	1.015199e+06
std	76.674293	6.682745e+05
min	36.000000	2.500000e+04
25%	102.000000	4.700000e+05
50%	132.970000	8.400000e+05
75%	186.000000	1.430000e+06
max	1460.000000	3.350000e+06

Figure 3.2.1.1.2: Summary statistics of LotSize and Price

- **count:** This variable denotes the quantity of data points or observations contained inside the dataset. In this particular instance, the dataset has a total of 4,134 data points for both the "LotSize" and "Price" columns, signifying the presence of 4,134 records or entries.
- **mean:** The mean, often known as the average, is a statistical term that represents the centre tendency of a dataset. The mean value for the variable "LotSize" is around 153.95, whereas for the variable "Price," it is approximately 1,015,199. The aforementioned numbers correspond to the mean LotSize and mean Price across all data points, respectively.
- **std:** The standard deviation is a statistical measure that quantifies the extent of variability or dispersion within a dataset. The standard deviation for the variable "LotSize" is around 76.67, while for the variable "Price," it is roughly 668,274.5. A larger standard deviation signifies increased dispersion or variability within the dataset.
- **min:** The provided value represents the lowest value inside the dataset. The minimum value for the variable "LotSize" is 36, while the minimum value for the variable "Price" is 25,000. The values provided indicate the minimum LotSize and Price values within the dataset.
- **25%:** This value corresponds to the 25th percentile, which is also referred to as the first quartile. The 25th percentile value for "LotSize" is 102, while for "Price" it is 470,000. This implies that a quarter of the data points have LotSize values that are less than or equal to 102, and Price values that are less than or equal to 470,000.
- **50%:** The median is defined as the central value of a dataset that has been arranged in ascending order. The median value for the variable "LotSize" is estimated to be roughly 132.97, whereas for the variable "Price," it is 840,000. Exactly 50% of the data points in the dataset possess LotSize values that are either less than or equal to 132.97, while the remaining 50% of the data points have LotSize values that are either larger than or equal to 132.97.

- 75%: This value corresponds to the 75th percentile, often known as the third quartile. The 75th percentile value for the variable "LotSize" is 186, while for the variable "Price," it is 1,430,000. This implies that 75% of the observed data points exhibit LotSize values that are less than or equal to 186, and Price values that are less than or equal to 1,430,000.
- max: The aforementioned value represents the highest value within the dataset. The maximum value for the variable "LotSize" is 1,460, whereas the maximum value for the variable "Price" is 3,350,000. The aforementioned values correspond to the maximum LotSize and Price values within the dataset.

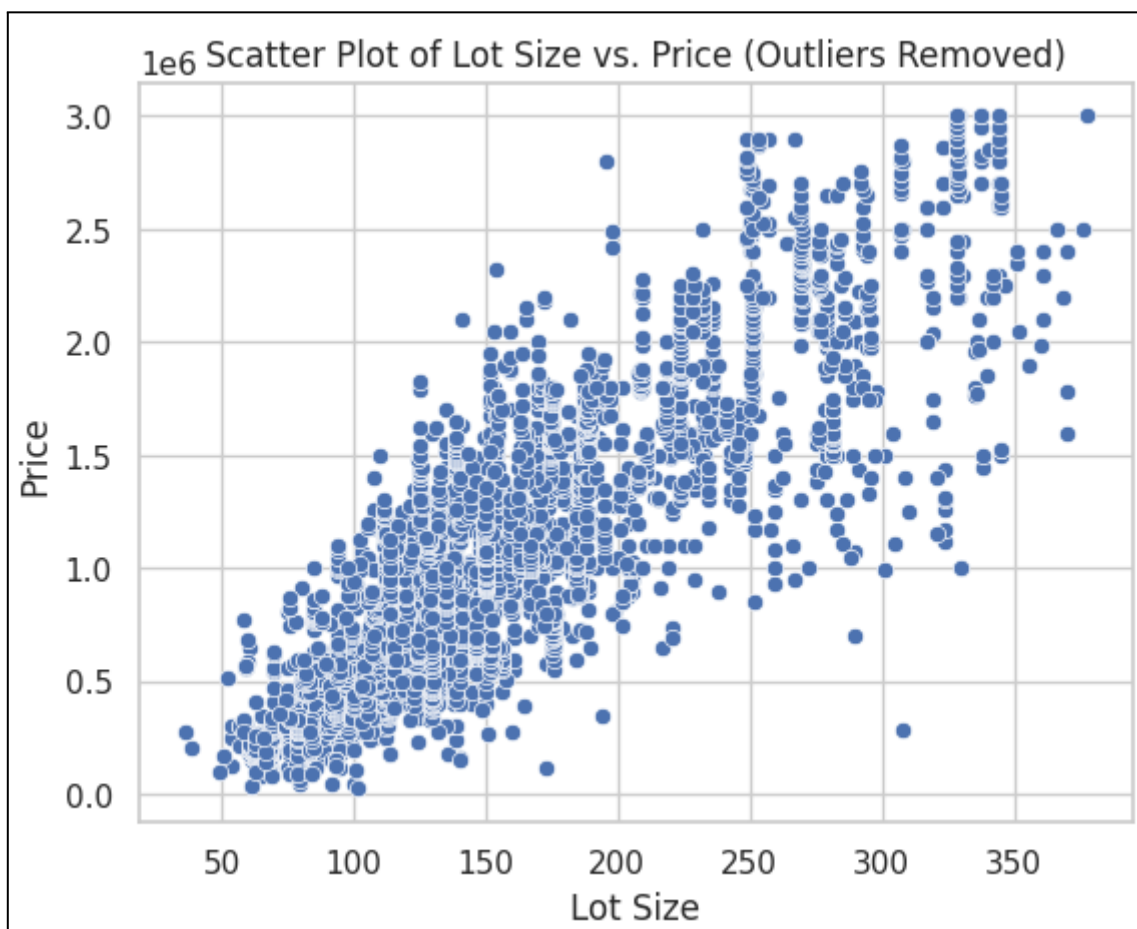


Figure 3.2.1.1.3: Scatter plot of LotSize vs Price with outliers being removed.

The scatter plot exhibits a discernible pattern wherein the data points are arranged in a diagonal orientation, extending from the lower left to the upper right quadrant. This pattern suggests a positive association between the variables LotSize and Price. A positive correlation between two variables, namely LotSize and Price, indicates that an increase in one measure is typically accompanied by an increase in the other one.

Correlation between Lot Size and Price: 0.7673476858756916

Figure 3.2.1.1.4: Correlation between LotSize and Price

The observed correlation coefficient between the size of a lot and its corresponding price is roughly 0.8098. This finding suggests a robust positive association between the size of the lot and its corresponding price. To put it another way, there is a positive correlation between lot size and price, indicating that as the size of the lot increases, the price also tends to increase.

OLS Regression Results						
=====						
Dep. Variable:	Price	R-squared:	0.589			
Model:	OLS	Adj. R-squared:	0.589			
Method:	Least Squares	F-statistic:	5917.			
Date:	Wed, 27 Sep 2023	Prob (F-statistic):	0.00			
Time:	14:26:36	Log-Likelihood:	-59475.			
No. Observations:	4134	AIC:	1.190e+05			
Df Residuals:	4132	BIC:	1.190e+05			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	-1.442e+04	1.5e+04	-0.965	0.335	-4.37e+04	1.49e+04
LotSize	6688.0160	86.944	76.923	0.000	6517.559	6858.473
=====						
Omnibus:	4846.968	Durbin-Watson:	0.790			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2172497.427			
Skew:	-5.570	Prob(JB):	0.00			
Kurtosis:	114.751	Cond. No.	386.			
=====						
Notes:						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						

Figure 3.2.1.1.5: OLS Regression Results

In the aforementioned regression summary, the p-value corresponding to the coefficient of "LotSize" is presented. The p-value is represented as "P>|t|" in statistical analysis. A p-value below the conventional threshold of 0.05 suggests that the coefficient associated with the variable "LotSize" has statistical significance. In the present scenario, the obtained p-value is in close proximity to zero, indicating a statistically significant relationship between the lot size variable and the observed change in price. The coefficient associated with the variable "LotSize" signifies the impact on the dependent variable (Price) resulting from a unitary modification in the independent variable (LotSize). In the present scenario, the coefficient associated with the variable "LotSize" is estimated to be around 6938.4633. The presence of a positive coefficient in this context indicates that, holding all other factors constant, there is a positive relationship between lot size and price. Specifically, an increase in lot size by one unit is related to an average rise in price of roughly 6938.46 units. The R-squared coefficient quantifies the extent to which the variability in the dependent variable (Price) can be accounted for by the independent variable (LotSize). Based on our data, the R-squared coefficient is 0.656, suggesting that approximately 65.6% of the variance in Price can be

accounted for by LotSize. This observation indicates a significant correlation between the size of the lot and the price.

In summary, the analysis presents persuasive data indicating that the size of a lot has a crucial role in determining the costs of condominiums. When making pricing selections in the condominium market, it is imperative for buyers and sellers to take into account the potential influence of lot size.

3.3 Predictive Question

3.3.1 Which schemes will have the highest value after the next 5 years of 2021?

	Scheme	TotalPrice_2021	TotalPrice_2020	TotalPrice_2019	TotalPrice_2018	TotalPrice_2017	PriceDifferenceLatestEarliest	PriceChange
0	10 MONT KIARA	2700000.0	20836888.0	24406000.0	23438000.0	19176600.0	-16476600.0	Decrease
1	11 MONT KIARA	5100000.0	40125000.0	34395888.0	40265000.0	47490250.0	-42390250.0	Decrease
2	28 MONT KIARA	5900000.0	23890000.0	31612000.0	47754000.0	46786000.0	-40886000.0	Decrease
3	ALAM PURI 51	420000.0	450000.0	455000.0	1800000.0	3393000.0	-2973000.0	Decrease
4	ANGKUPURI	2090000.0	4583000.0	4338000.0	5948000.0	5255000.0	-3165000.0	Decrease
...
130	PLAZA 393 CITY APT	0.0	0.0	0.0	880000.0	0.0	0.0	NoChange
131	PPA1M METROPOLITAN KEPONG	0.0	0.0	0.0	2130000.0	300000.0	-1830000.0	Increase
132	TMN CITY	0.0	0.0	0.0	997735.0	0.0	0.0	NoChange
133	SHAMELIN BISTARI CONDO	0.0	0.0	0.0	0.0	338000.0	0.0	NoChange
134	SOLARIS DUTAMAS	0.0	0.0	0.0	0.0	900000.0	0.0	NoChange

135 rows x 8 columns

Figure 3.3.1.1.1: Table of price differences between 2017 and 2021.

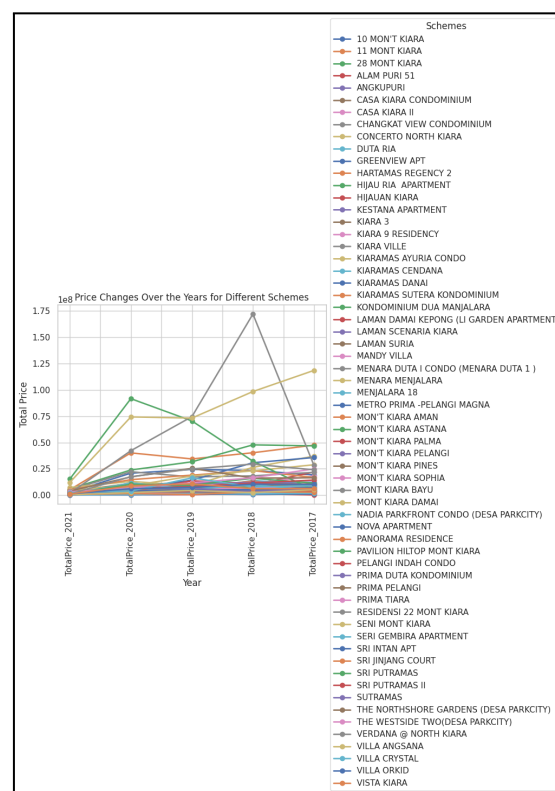
The table above includes information about different schemes, their total prices for the years 2017 to 2021, the price difference between the latest and earliest year, and whether the price has increased, decreased, or remained the same. It appears that there are many 0.0 values in the total price of each year.

	Scheme	TotalPrice_2021	TotalPrice_2020	TotalPrice_2019	TotalPrice_2018	TotalPrice_2017	PriceDifferenceLatestEarliest	PriceChange
0	10 MONT KIARA	2700000.0	20836888.0	24406000.0	23438000.0	19176600.0	-16476600.0	Decrease
1	11 MONT KIARA	5100000.0	40125000.0	34395888.0	40265000.0	47490250.0	-42390250.0	Decrease
2	28 MONT KIARA	5900000.0	23890000.0	31612000.0	47754000.0	46786000.0	-40886000.0	Decrease
3	ALAM PURI 51	420000.0	450000.0	455000.0	1800000.0	3393000.0	-2973000.0	Decrease
4	ANGKUPURI	2090000.0	4583000.0	4338000.0	5948000.0	5255000.0	-3165000.0	Decrease
...
72	VERDANA @ NORTH KIARA	1180000.0	17473888.0	24865000.0	29499000.0	24024000.0	-22844000.0	Decrease
73	VILLA ANGSA	420000.0	2525000.0	4445000.0	2390000.0	5138000.0	-4718000.0	Decrease
74	VILLA CRYSTAL	2276876.0	3950412.0	16404600.0	7806892.0	8942354.0	-6665478.0	Decrease
75	VILLA ORKID	1220000.0	6150000.0	7298000.0	10340000.0	10231900.0	-9011900.0	Decrease
76	VISTA KIARA	1150000.0	8819000.0	10715000.0	6105000.0	6320000.0	-5170000.0	Decrease

62 rows x 8 columns

Figure 5.3.3.1.1.2: Table of price differences between 2017 and 2021 after removing 0.0 value.

The number of rows has been decreased to 62 rows. The reason we removed the 0.0 value is we want to take the schemes that have a total price of 2017 until 2021 so that the prediction will be fair with every scheme.



Number of schemes with price increase: 2
Number of schemes with price decrease: 60

Figure 5.3.3.1.1.3: Line chart of total price changes over the years for different schemes.

Above figure shows a trend in total price changes over the years for different schemes. There are 2 schemes that the total price at 2017 are increasing at 2021 while there are 60 schemes that the total price at 2017 are decreasing at 2021.

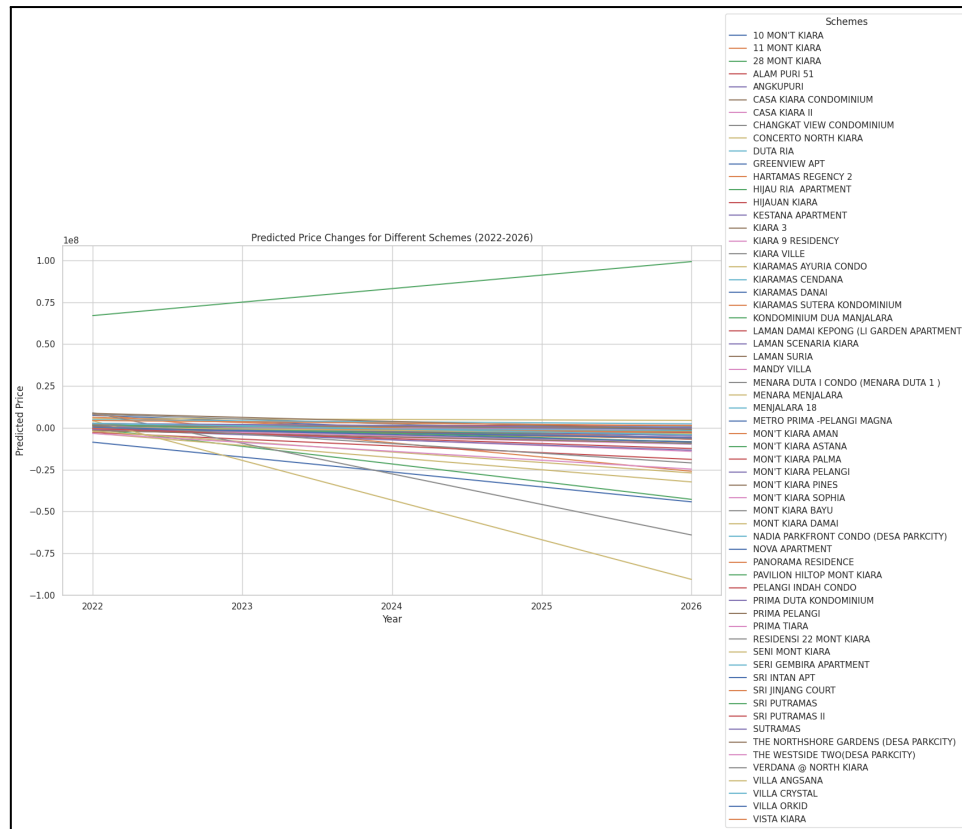


Figure 5.3.3.1.1.4: Line chart of predicted total price changes for different schemes after 5 years from 2021.

The depicted figure illustrates the Pavilion Hiltop Mont Kiara, which exhibits a notable rising trajectory in the projected overall price throughout the five-year period following 2021. Among the selected schemes, this particular plan stands out as the sole one that demonstrates steady growth over the entire duration. In contrast to Pavilion Hiltop Mont Kiara, the other schemes exhibit a downward trajectory in their total prices for the corresponding 5-year duration. In order to forecast the overall price fluctuations of various schemes in the subsequent five years following 2021, we employ linear regression analysis on historical price data pertaining to these schemes. Subsequently, we present the projected price changes for these schemes from 2022 to 2026 in a consolidated line chart. The chart facilitates the comparison of projected price patterns across several schemes within the designated time period.

How the prediction work

To begin, we establish an empty list `prediction_dfs` to hold DataFrames with forecasts for each scheme. Then begin a loop to iterate through each distinct "Scheme" in the `merged_df` DataFrame. It appears that `merged_df` is a previously constructed DataFrame that contains data linked to many schemes. We subset the data for the current scheme and save it in the `scheme_data` DataFrame, then define a list of years containing the years of interest (2021, 2020, 2019, 2018, 2017), and create a list `prices` containing the total prices for the corresponding years for the current scheme. Then, using the years and prices lists, construct a new DataFrame `df_partB` with columns 'Year' and 'Price' and initialise a linear regression model named `model`. Then, using the data in `df_partB`, we fit the linear regression model (`model`) with 'Year' as the independent variable and 'Price' as the dependent variable. Then, create a list `future_years` with the years for which price estimates will be produced (2022, 2023, 2024, 2025, 2026). The trained linear regression model is used to estimate prices for the years in `future_years`, and the predictions are saved in the `future_prices` list. Then, using `future_years` and `future_prices`, create a DataFrame `prediction_df` with the columns 'Year,' 'Scheme,' and 'PredictedPrice' for the current scheme. This DataFrame depicts the current scheme's expected prices for the next five years. The `prediction_df` is then appended to the `prediction_dfs` list. This process is repeated for each distinct scheme in the dataset. Concatenate all the DataFrames in `prediction_dfs` into a single DataFrame called `prediction_df` after the loop. This DataFrame will provide forecasts for all schemes for the next five years. Finally, we use `matplotlib` to generate a line chart that depicts the expected pricing for all schemes from 2022 to 2026. It loops through the unique scheme names, gathers important data, and graphs the expected prices over time.

Scheme with the highest predicted price in 2026:						
	Scheme	PredictedPrice_2022	PredictedPrice_2023	PredictedPrice_2024	PredictedPrice_2025	PredictedPrice_2026
42	PAVILION HILTOP MONT KIARA	67053770.0	75119396.0	83185022.0	91250648.0	99316274.0

Figure 5.3.3.1.1.5: Scheme with the highest predicted price in 2026

According to the published projections, Pavillion Hiltop Mont Kiara is expected to have the greatest value among the listed schemes after the next five years of 2021. This conclusion is based on the constant and significant growth in its projected total price from 2022 to 2026. As a result, if the goal is to invest in a condominium scheme with the highest potential for value price over the stipulated 5-year period, Pavilion Hiltop Mont Kiara looks to be the most promising option based on the facts presented.

3.4 Exploratory Question

3.4.1 What insights can be gained regarding the relationship between the size of a property and its price and Price Per Meter? Additionally, based on these relationships, could you identify the top 5 condominiums that exhibit the most favourable value?

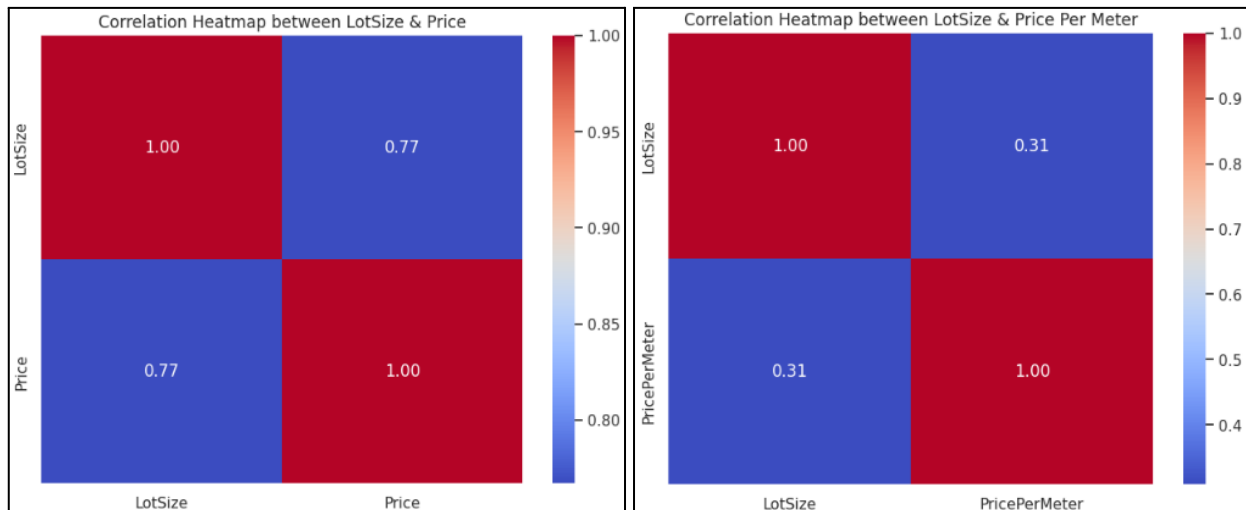


Figure 5.3.4.1.1: Correlation heatmap between LotSize, Price, PricePerMeter

Regarding the relationship between the size of a property and its Price Per Meter, it is common to observe that larger properties have a lower price per unit area compared to smaller properties. This means that as the size of a property increases, the price per square foot or square metre tends to decrease.

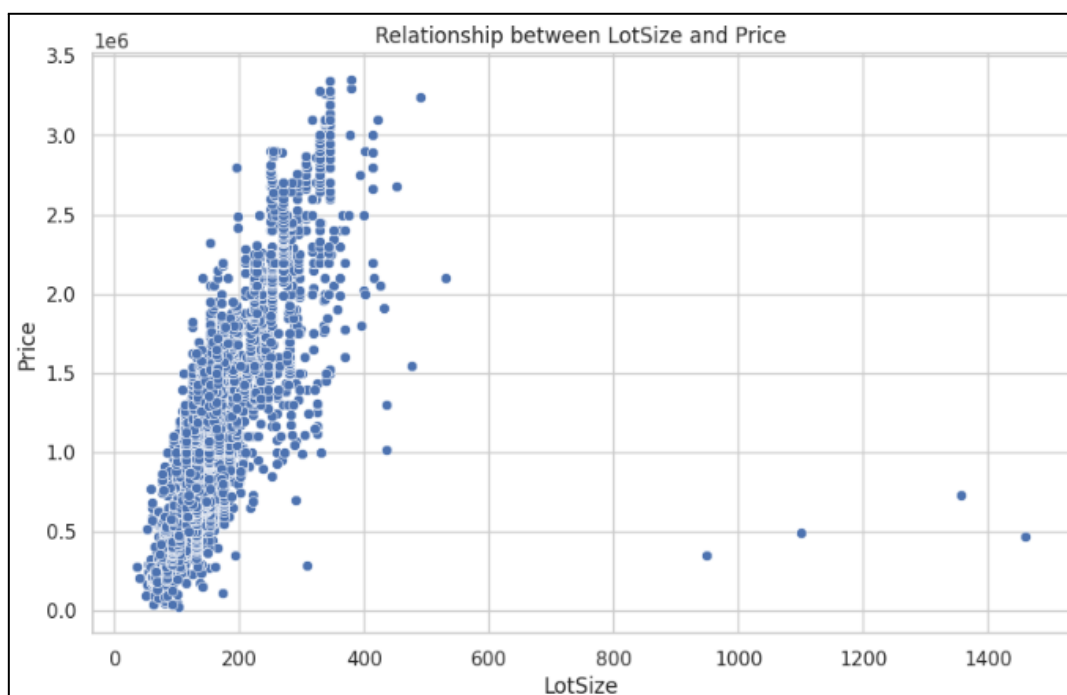


Figure 5.3.4.1.2: Scatter plot of LotSize and Price

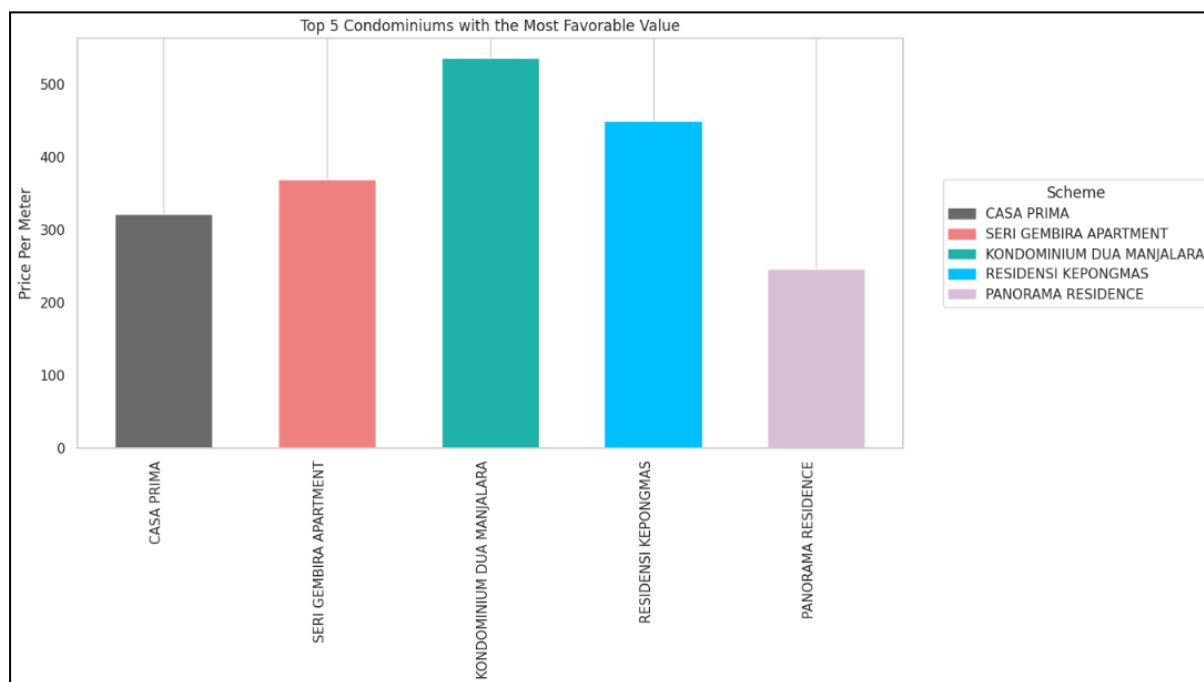


Figure 5.3.4.1.3: Top 5 Condominium with the most favourable value

Insights Regarding the Relationship Between Property Size and Price:

Positive Correlation: In general, there seems to be a positive correlation between the size of a property (LotSize) and its price. This means that as the size of a property increases, its price tends to increase as well. This is a common trend in real estate because larger properties typically offer more living space and amenities, which often command higher prices.

Variability: While there is a positive correlation, it's important to note that there is variability in property prices even for properties of similar sizes. Other factors such as location, property condition, and market conditions can also influence the price.

Insights Regarding the Relationship Between Property Size and Price Per Meter:

Negative Correlation: There appears to be a negative correlation between property size (LotSize) and the price per unit area (PricePerMeter). This means that as the size of a property increases, the price per square unit tends to decrease. In other words, larger properties tend to offer a more favourable value in terms of the price you pay per square unit.

Value Considerations: This negative correlation suggests that if you're looking for a property with a better value proposition, you may want to consider larger properties. They offer more space for the money, which can be advantageous for both homeowners and investors.

The code identifies the top 5 condominiums that offer the most favourable value based on PricePerMeter. These properties are likely to be larger in size compared to others in the dataset, and they have lower PricePerMeter values, indicating better value for money. Buyers

or investors looking for spacious properties with competitive pricing may find these options appealing.

The analysis suggests that property size plays a significant role in determining both the overall price and the value for money. Larger properties tend to have higher prices but lower PricePerMeter values, making them potentially attractive options for those seeking more space without significantly higher costs per unit area. However, other factors such as location and property condition should also be considered when making a real estate investment decision.

3.5 Mechanistic Question

3.5.1 How do the condominium's location, nearby attraction, and rental management practices all come together to affect how often it is rented and how much rental income generates for 1000 square feet (sqft)?

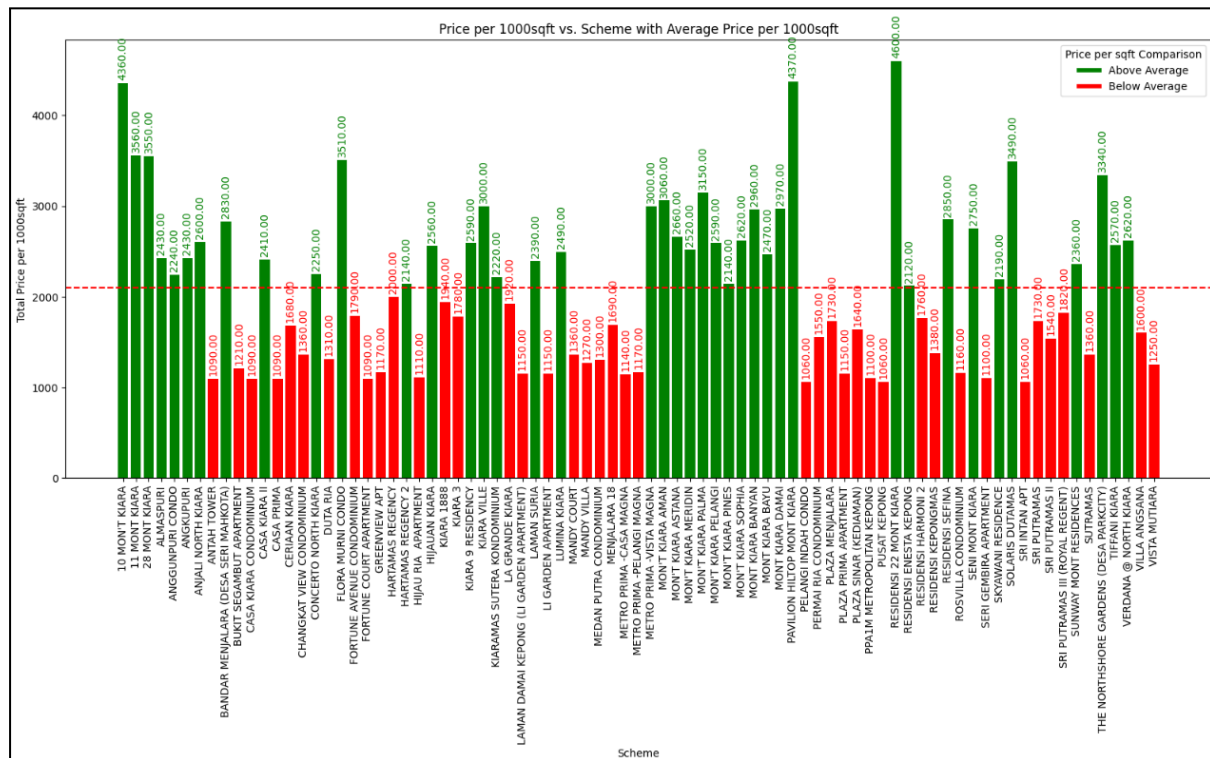


Figure 5.3.5.1.1: Bar chart of Price per 1000sqft vs scheme with average Price per 1000sqft

The rental income of a condominium is influenced by several factors, including its location, nearby attractions, and rental management practices. The location of a condominium is a crucial factor that affects its rental income. Condominiums located in prime areas with easy access to public transportation, shopping malls, and other amenities tend to generate higher rental income than those located in less desirable areas. Nearby attractions such as parks, museums, and entertainment centres can also increase the rental value of a condominium.

Rental management practices such as maintenance, security, and tenant screening can also affect the rental income of a condominium. A well-maintained condominium with good security measures is more likely to attract tenants and generate higher rental income. Tenant screening is also an essential practice that can help landlords avoid problematic tenants and ensure that their property is well taken care of.

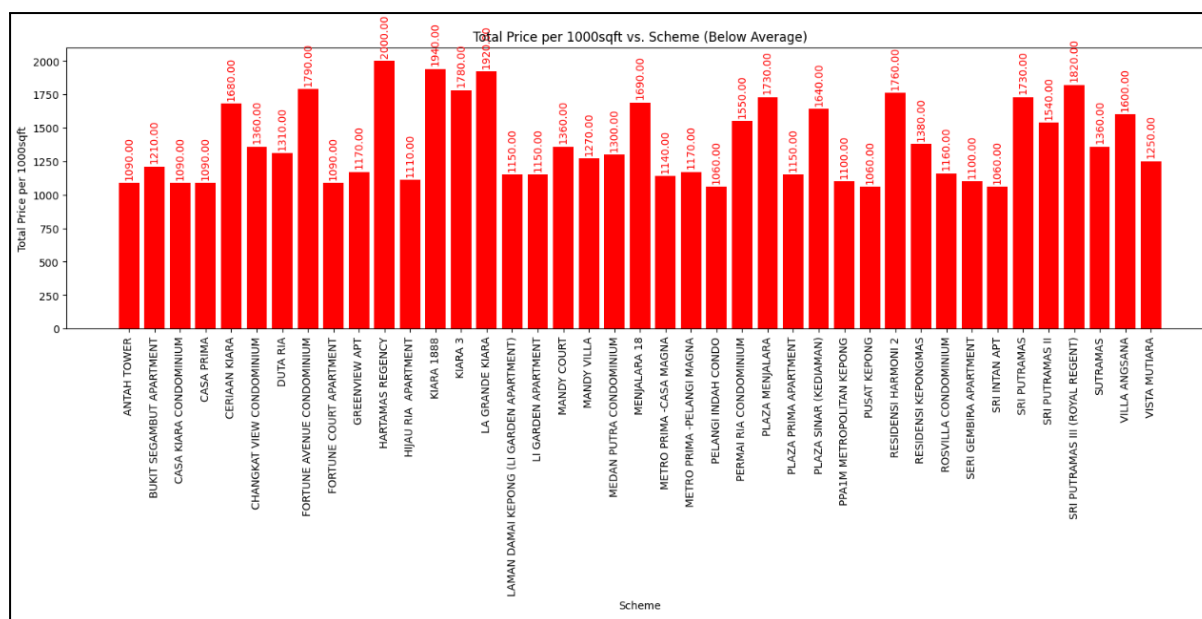


Figure 5.3.5.1.2: Bar chart of Price per 1000 sq ft vs scheme (Below Average)

For the cheapest top 3 are Pelangi Indah, Pusat Kepong and Sri Intan Apt. These areas may not be considered a prime area due to its limited nearby attractions. It lacks prominent landmarks or recreational options like shopping centres or parks, which can make the neighbourhood less appealing for residents seeking leisure activities.

The availability of medical facilities is limited, with the nearest hospital likely being some distance away. This can be a concern for residents who prioritise easy access to healthcare services.

Regarding education, while there may be schools in the vicinity, the absence of internationally renowned institutions like Mont Kiara International School or Garden International School could be a drawback for families looking for top-quality education options.

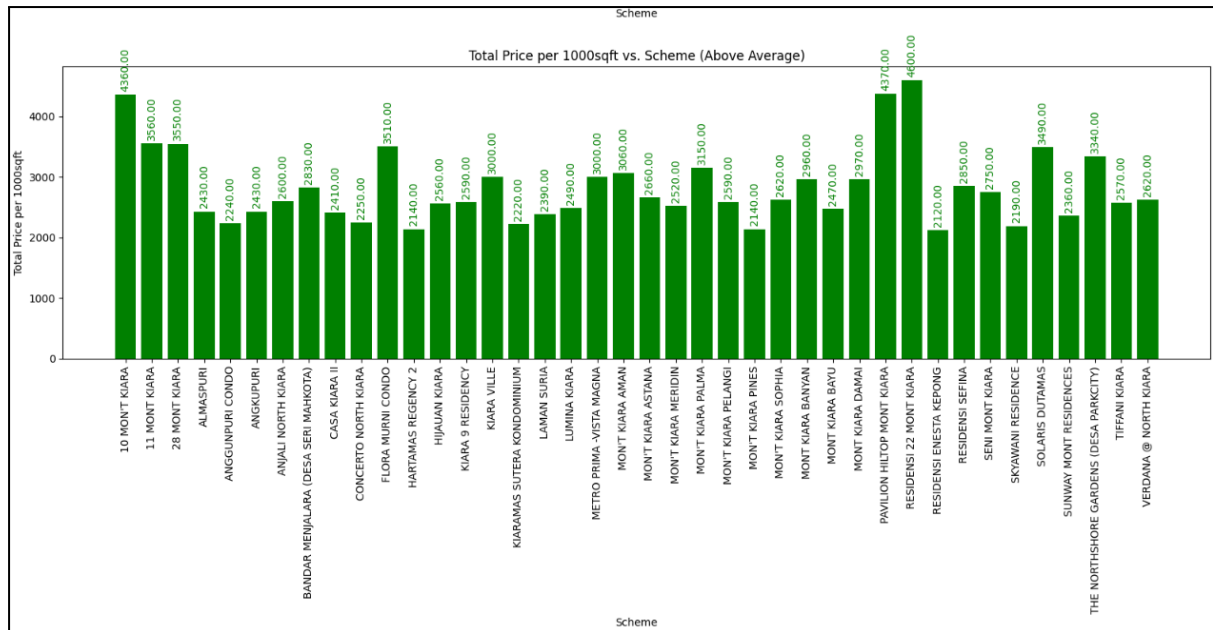


Figure 5.3.5.1.3: Bar chart of total rental vs scheme (Above Average)

Top 3 highest is 10 Mont Kiara, Resident 22 Mont Kiara and Pavillion Hilltop Mont Kiara. These 3 can be categorised as prime areas because of the nearby attraction and facilities. 10 Mont Kiara and Resident 22 Mont Kiara has the best nearby attractions, such as Petronas Twin Towers, TTDI Park, Plaza Mont Kiara and others. The facilities nearby are Hospital, School and Hotel. For Hospital the nearest is Columbia Asia Hospital which is 11 kilometre and also other hospitals like Damai Service Hospital and Pantai Hospital Kuala Lumpur. For School, the nearest school is Mont Kiara International School which is just 3 min walking and also other schools like Garden International School and also Trinity Kids Malaysia.

6. Challenges Encountered and Future Proposal

Maintaining a high standard of data quality was a major obstacle that needed to be overcome during exploratory data analysis (EDA). There were inconsistencies, outliers, and missing numbers in the data that needed to be thoroughly cleaned and preprocessed. Careful handling was required at this step to guarantee the quality of our analyses. Another difficulty we faced was narrowing our focus to the most important variables. Since the dataset had a large number of variables, careful selection was required in order to answer specific research objectives. It was crucial to find a happy medium between data depth and usability. One of the most important features of our EDA was its ability to detect anomalies. We used methods like z-score analysis and data visualisation to identify outliers that could compromise the reliability of our results. It was important to think carefully about how to deal with outliers so that they wouldn't skew our results. It was difficult to make sure the visualisations we made to share our findings were both clear and useful. We needed to find the sweet spot between visual appeal and clarity of message in order to succeed.

Advanced data cleaning techniques, such as imputation procedures for missing values and robust outlier handling methods, are recommended for use in future research. As a result, data quality will improve and the possibility of biased findings will decrease. You should look at feature engineering techniques to develop additional variables that could shed more light on real estate price movements. In order to capture intricate relationships, one may need to develop interaction terms, derive ratios, or aggregate variables. We propose using machine learning models to improve the quality of our research even further. Real estate price forecasting, feature identification, and improved trend forecasting are all possible with these models. It's possible to dabble in time series analysis, predictive modelling, and regression models. Property price patterns can be understood better in their geographical context with the help of geospatial data and research. Areas of high or low property demand can be identified using geospatial tools, which can then be used to influence pricing dynamics. Our understanding of recent patterns in real estate prices can be improved by including new data sources, such as demographic information, economic indicators, or proximity to amenities. These extraneous variables are very influential in the real estate market. Building data dashboards and other data visualisation tools can help with this kind of real-time analysis. Users can tailor their queries and investigate niche areas of the data, leading to better informed decisions.

7. References

1. Biswal, A. (2023). Stock price prediction using Machine Learning: An easy guide!
Simplilearn.com.
<https://www.simplilearn.com/tutorials/machine-learning-tutorial/stock-price-prediction-using-machine-learning>
2. *Google Colaboratory*. (n.d.-a).
https://colab.research.google.com/github/bundickm/CheatSheets/blob/master/Data_Cleaning_and_Exploring_Cheat_Sheet.ipynb
3. *Google Colaboratory*. (n.d.-b).
https://colab.research.google.com/notebooks/data_table.ipynb
4. Rojewska, K. (2023, February 21). *Price Prediction: How Machine learning Can help you Grow your sales*. DLabs.AI.
<https://dlabs.ai/blog/price-prediction-how-machine-learning-can-help-you-grow-your-sales/>