



DATA ANALYSIS PROJECT

REAL-ESTATE IN MELBOURNE, AUSTRALIA

STA1020-B

SUMMER SEMESTER

GROUP 3

30/07/2024

**Kyeremateng martin**

**Analysis of Real Estate dataset from Melbourne, Australia to identify patterns and make inference from the data. Using key attributes in the dataset, we have made analysis on how each key attribute impacts the market price of real estate properties in Melbourne.**

#### Q1.

The dataset originates from real estate transactions in Melbourne, Australia. It captures extensive details about properties sold in specific regions, with 13,580 observations in each attribute. It includes various attributes such as the street address, type of property (e.g., houses, units, townhouses), and suburb location. Additionally, the dataset records the method of sale, which cuts across a range of outcomes from sold to withdrawn or passed in. It has Key transactional details like the sale price, the real estate agent responsible for the sale, and the date of purchase which are are documented. Geographic context is provided through the distance of each property from the Central Business District (CBD) and the general region name, such as West or North East. The dataset also includes structural details such as the number of rooms, bathrooms, car spots, land, and building sizes. It further notes the total number of properties in each suburb and the governing council area. The dataset is well-suited for analyzing market trends, property valuations, and the impact of various factors on real estate prices (*Melbourne Housing Snapshot*, n.d.)

Key Variables

* Address: The street address of the property.
* Type: Categorizes the type of property. Possible values include:

h: House, cottage, villa, semi, terrace

u: Unit, duplex

t: Townhouse

* Suburb: The suburb where the property is located.
* Method: Indicates how the property was sold. Possible values include:

S: Sold

SP: Sold prior

PI: Passed in

PN: Sold prior not disclosed

SN: Sold not disclosed

NB: No bid

VB: Vendor bid

W: Withdrawn prior to auction

SA: Sold after auction

SS: Sold after auction price not disclosed

N/A: Price or highest bid not available

* Rooms: Number of rooms in the property.
* Price: Sale price of the property in dollars.
* Real Estate Agent (SellerG): The agent or agency responsible for the sale.
* Date of Sale (Date): The date on which the property was sold.
* Distance from C.B.D. (Distance): The distance of the property from the Central Business District (CBD).
* Region Name (Regionname): General region of the property, such as West, North West, North, North East.
* Property Count (Propertycount): The total number of properties in the suburb.
* Number of Bedrooms (Bedroom2): Number of bedrooms obtained from an additional data source.
* Number of Bathrooms (Bathroom): Number of bathrooms in the property.
* Number of Carspots (Car): Number of car parking spaces available with the property.
* Land Size (Landsize): The size of the land on which the property is situated.
* Building Size (BuildingArea): The size of the building area of the property.
* Governing Council Area (CouncilArea): The local governing council for the area where the property is located.

#### Q2.

Research Question

What is the impact of Real Estate key attributes on housing prices in Melbourne.

Specific Questions Under Analysis

1. Which region has the most expensive real estate market?
2. Which property type has the highest market price?
3. Is there any relationship between Building Area and market price?
4. Does the location of a property determine it’s price?

#### Q3.

Methodology

Data was first downloaded as a .csv file from [https://www.kaggle.com/datasets](https://www.kaggle./datasets) which was shared to group members. A project was created where individual members were given the authorization to work on independently. Using Jupyter Notebook, the data was loaded into a python DataFrame and a wrangle function created to preprocess and clean the data. The necessary libraries were imported and data explored to see key attributes that are suitable for the analysis. Background research was done to know the range of values that a typical real estate data and its attributes should have in order to be able to identify outliers and clean the data. We cleaned the data by handling missing values using SimpleImputer library, normalized some distributions for analysis, and attributes which were not of greater need to our analysis dropped successfully. A target vector/variable [Price\_aud] was chosen for analysis based on the research question to determine how each feature (categorical/numerical) affects the target vector. The target {Price\_aud] was in Australia dollar and we converted to United States dollar to create a general market price that could easily be understood by even the layman (Shukla, 2022)

Key Variables Chosen for Analysis

Categorical Variables

* Type
* Region Name

Numerical Continuous Variables

* Price
* Latitude
* Longitude
* Building Area
* Distance
* Rooms

**Research Question 1**

Which region in Melbourne has the most expensive real estate market?

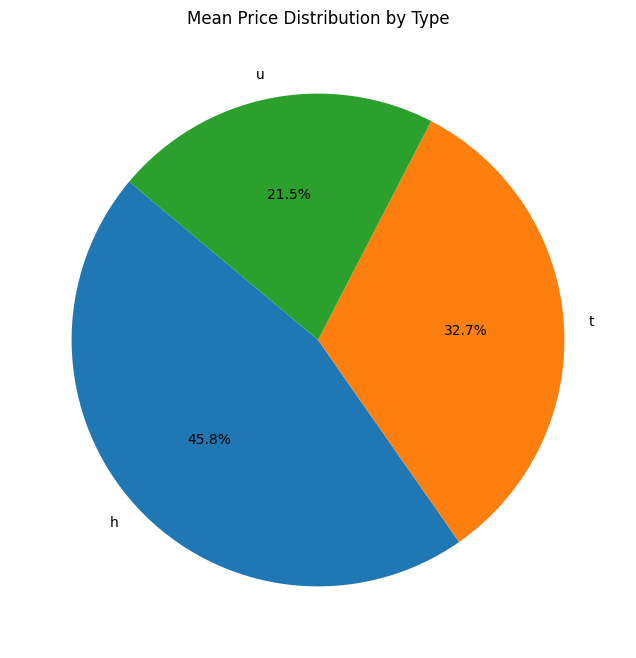


|  |  |  |
| --- | --- | --- |
| Region Name | Mean Market Price [USD] | Percentage Mean |
| Southern Metropolitan | 1.043705e+06, 1,043,705.00 | 20.04% |
| Eastern Metropolitan | 8.265258e+05, 826,525.80 | 16.2% |
| South-Eastern Metropolitan | 6.950531e+05, 695,053.10 | 13.6% |
| Northern Metropolitan | 6.500155e+05, 650,015.50 | 12.7% |
| Western Metropolitan | 6.460136e+05, 646,013.60 | 12.6% |
| Eastern Victoria | 5.096376e+05, 509,637.60 | 10.0% |
| Northern Victoria | 4.442643e+05, 444,264.30 | 8.7% |
| Western Victoria | 3.017438e+05, 301,743.80 | 5.9% |

According to the analysis presented, the region with the most expensive real estate market in Melbourne, Australia is Southern Metropolitan, with a mean market price of $1,043,705.00.

**Research Question 2.**

Which property type in Melbourne has the most expensive market price?

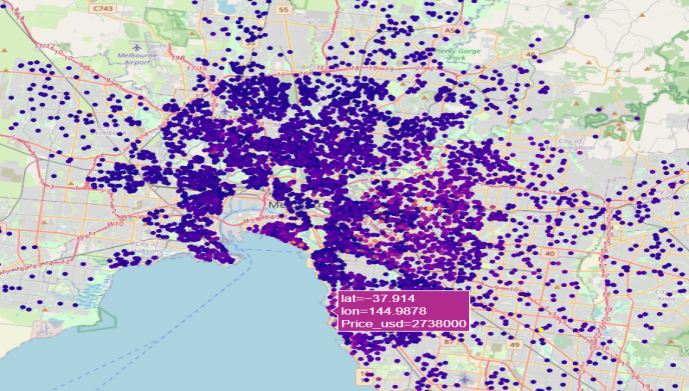
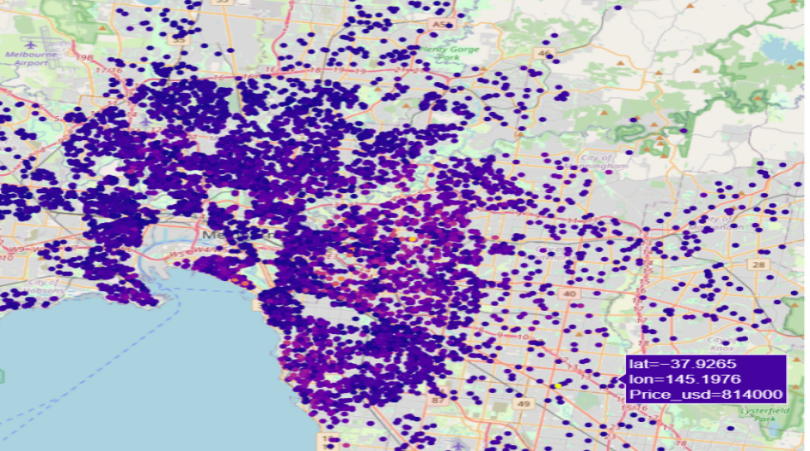
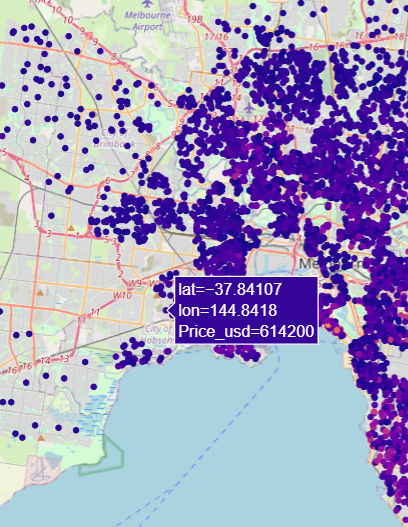
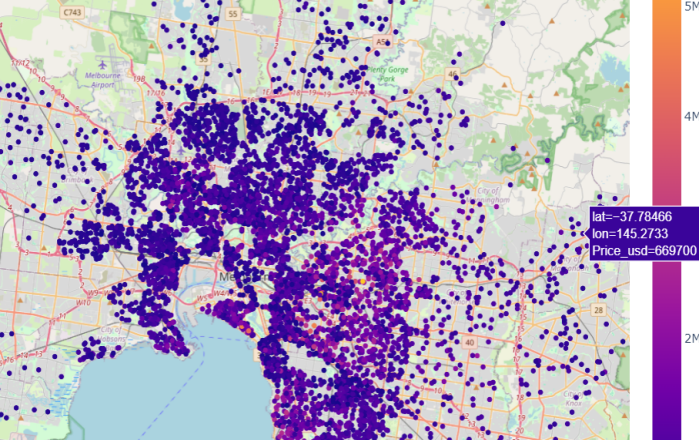


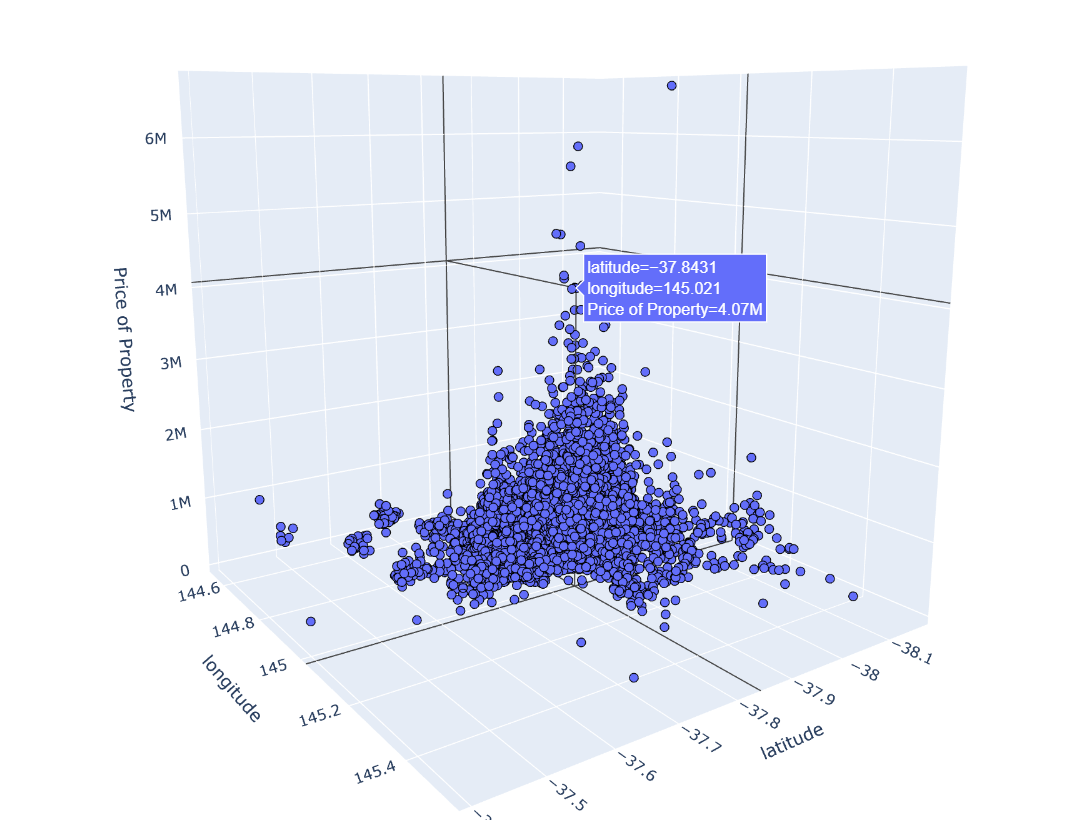
This graph shows that market price for property type h - house,cottage,villa, semi,terrace in Melbourne has the most expensive market price with a mean market price of $ 935,245.957169

|  |  |  |
| --- | --- | --- |
| Property Type | Mean Market Price [USD] | Percentage Mean |
| h - house,cottage,villa, semi,terrace | 935,245.957169 | 45.8% |
| t - townhouse; developing site, other residential | 667,507.453353 | 32.7% |
| u - unit, duplex. | 438,562.387967 | 21.5% |

**Research Question 3**

Is there any relationship between Location and market price?

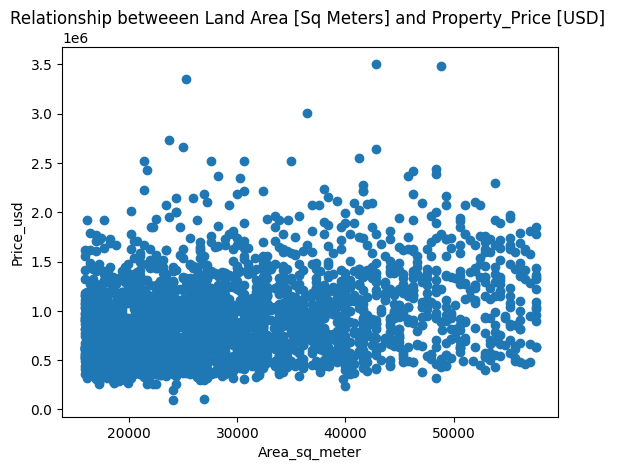
From the figures, we could possibly infer that there is a relationship between location and price of property. Properties that are within the Southern Metropolitan seem to have higher property prices, followed by those in the Eastern metropolitan and it follows as described in the bar chart previously.



3-D visualization of location impact on property price..

**Research Question 4.**

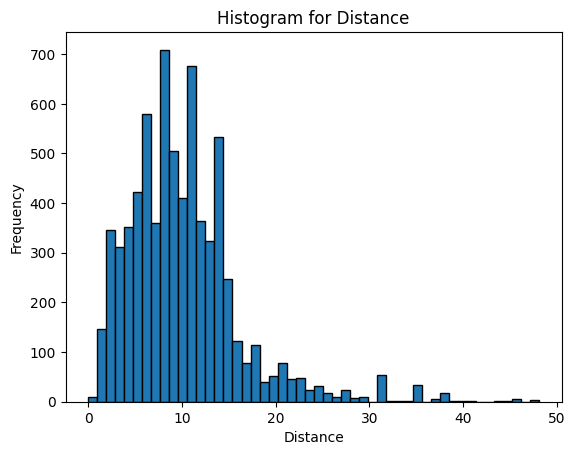
Impact of Building area on property type.



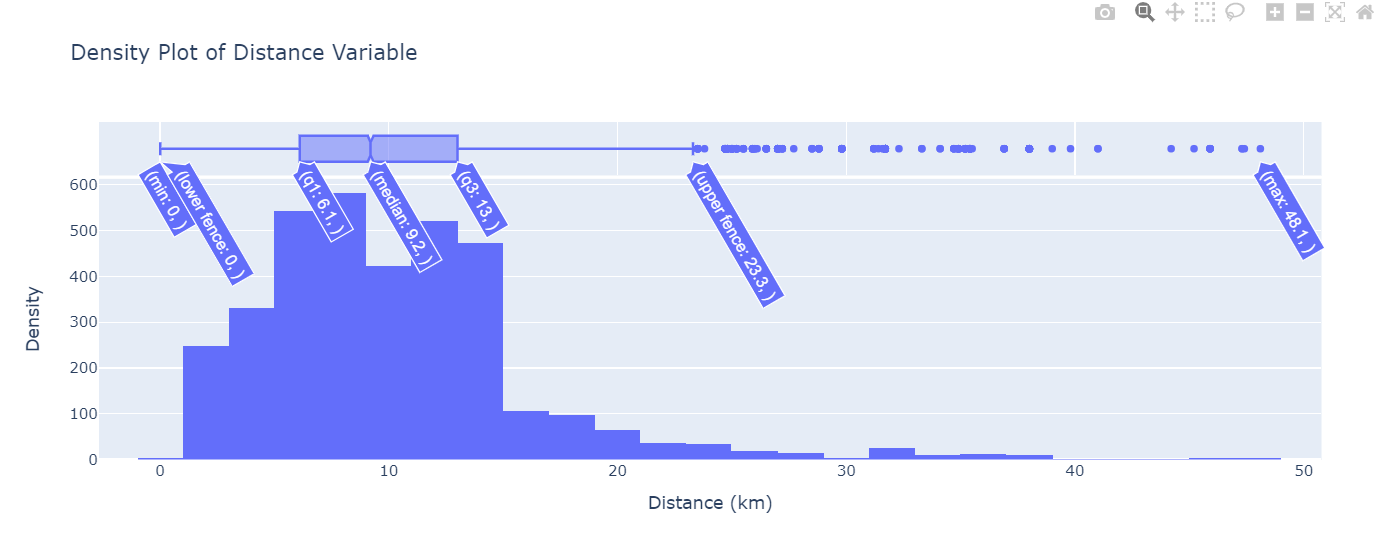
Correlation Coefficient (method=’spearman’) = 0.3054862861702415

This scatter plot shows the relationship between Land area and property market price. This figure shows that, there is a weak correlation between land area/ building area and property price. This could be because there are so many driving factors that affect the price of a property especially in Melbourne. Research has shown that properties near the ocean and those near the forest regardless of their size has higher prices than those distant from the sea. In-addition, property type, distance from CBD, region and other factors account greatly for the price of a property. Building area/land area in this case could be a driving factor since there’s a kind of weak correlation but cannot account greatly as a determinant for pricing.

**Histogram for distance (km) from CBD**



The distribution shows a slight skewness which is positive. This means there is a possibility of large values of outliers (Long distance) making the distribution skewed to the right.



Box-plot showing distribution of property distance from CBD. The figure indicates that there are outliers above the maximum, the whisker between the upper quartile and the maximum is longer than that between the minimum and the lower quartile indicating large spread of datapoints above the median. The minimum of the values in distance is 0.00 which is an outlier since there is no zero distance. The maximum is 48.1km, median 9.2km, third quartile 13.0km, first quartile 6.1km.

Spread and Skewness: The significant difference between the maximum (48.1 km) and the third quartile (13.0 km) suggests that the data is right-skewed, with a long tail of higher values. The median being closer to Q1 than Q3 are evidence supporting this skewness.

Central Tendency: The median (9.2 km) gives a good central value for the dataset, but given the skewness and the presence of outliers, the mean might be higher than the median.

Variability: The IQR of 6.9 km shows a moderate spread of the central 50% of the data. However, the presence of outliers (especially the zero distance) and the large maximum value indicates high variability in the dataset.

Outliers and Data Quality: The zero distance outlier indicates a potential issue with data quality or collection.

#### Q4.

Continuous Variable : Building Area

Descriptive Statistics of Building Area in [square\_meters]

count 2867.000000

mean 2875.607189

std 1080.624525

min 1587.600000

Mode: 1690.0

Median: 2560.0

Variance: 11749.363

max 5760.000000

Range= max-min= (5760.000 - 1587.600)= 4172.40 square\_meters

Interpretation of Results

The dataset comprises 2,867 observations, providing adequate sample for understanding the characteristics of building areas in the real estate market in Melbourne.

Central Tendency

The mean building area is approximately 2,875.61 square meters. This figure represents the average size of properties within the dataset and provides a general sense of the typical property size. However, the mean value is influenced by the presence of some large values (outliers), which skews the average upwards. In contrast, the median building area is 2,560 square meters. As the median represents the middle value when the data is ordered, it is less sensitive to skewed data than the mean. The fact that the median is lower than the mean suggests that the distribution of building areas is right-skewed. This right skew indicates that while most properties are clustered around the median value, there are a considerable number of larger properties that extend the distribution to the right.

Dispersion and Variability

The standard deviation of 1,080.62 square meters and variance of 11,749.36 square meters indicate a high degree of variability in building sizes. These measures highlight the spread of property sizes around the mean. A high standard deviation of 1080.624 sq\_m implies that building areas deviate significantly from the average size, suggesting a diverse range of property sizes in the market. The large variance further shows that the data encompasses a wide range of building areas, from smaller to much larger properties.

Extreme Values

The minimum building area recorded is 1,587.60 square meters, while the maximum reaches 5,760 square meters. This substantial range of 4,172.4 square meters illustrates the extent of variability within the dataset. The presence of such extreme values in real estate data in Melbourne as researched do not always constitute outliers but reflect the diverse nature of the Melbourne real estate market.

Most Common Building Size

The mode, which is the most frequently occurring building area, is 1,690 square meters. This mode represents the building size that appears most often in the dataset, highlighting a common property size. Despite this, the mode is considerably lower than both the mean and the median, which suggests that while 1,690 square meters is a common size, it is not as representative of the overall market distribution, which is skewed towards larger properties.

#### Q5.

Variable: Building Area

This variable has 2560 observations. Creating and interpreting stem and leaf plot for this data is not feasible so we did aggregation to be able to fit all observations representing the different observations in the BuildingArea variable in chunks to be able to interpret.

Range 1587.6-2087.6:

Range 2087.6-2587.6:

Range 2587.6-3087.6:

Range 3087.6-3587.6:

Range 3587.6-4087.6:

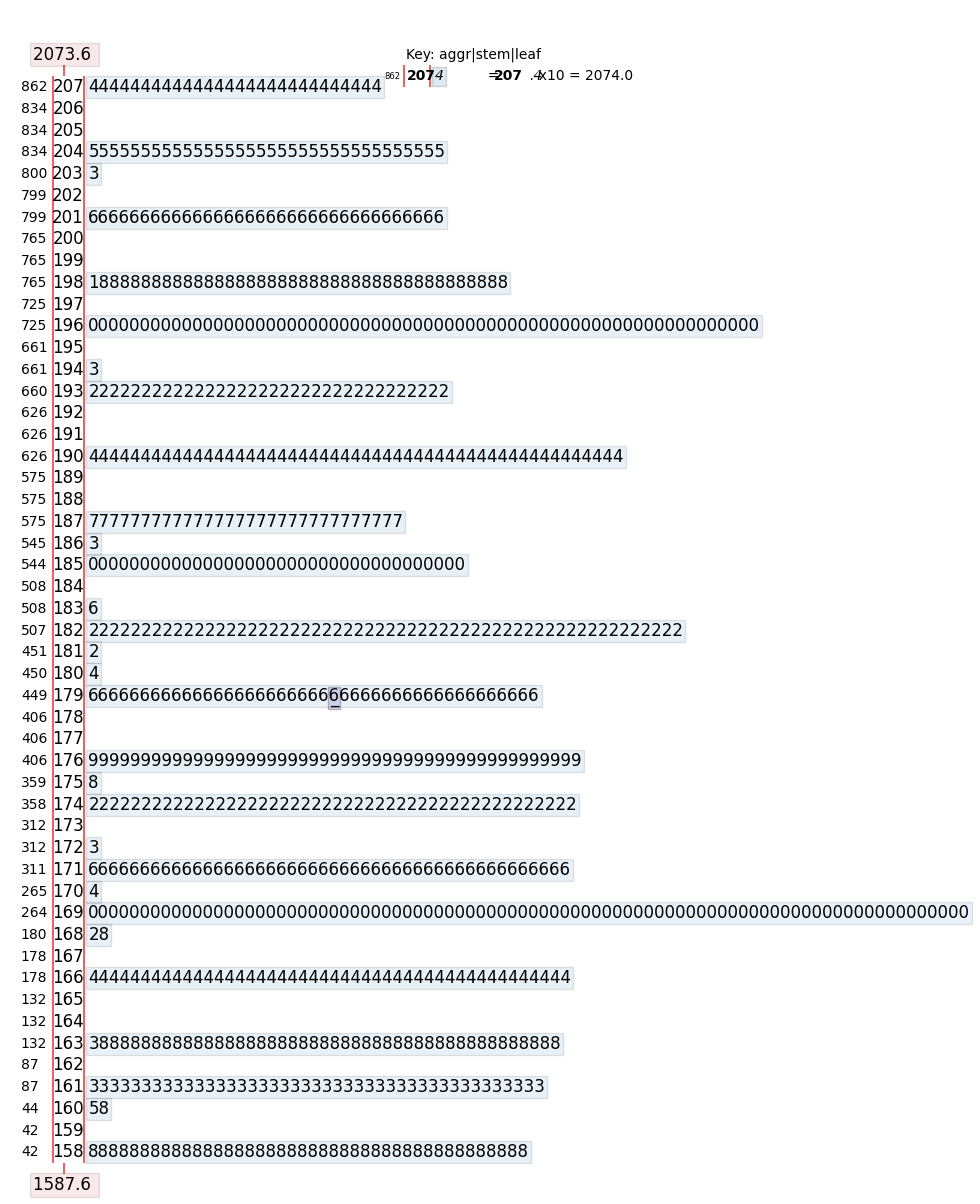
Range 4087.6-4587.6:

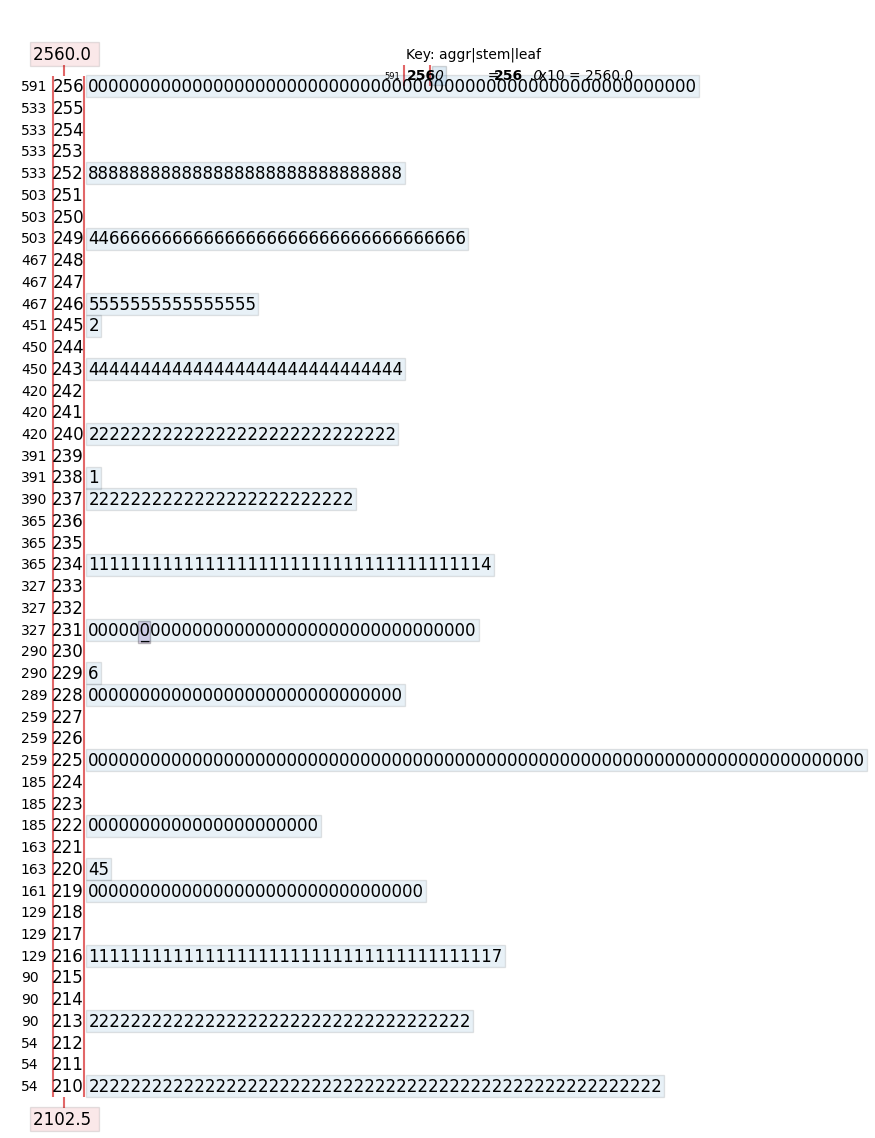
Range 4587.6-5087.6:

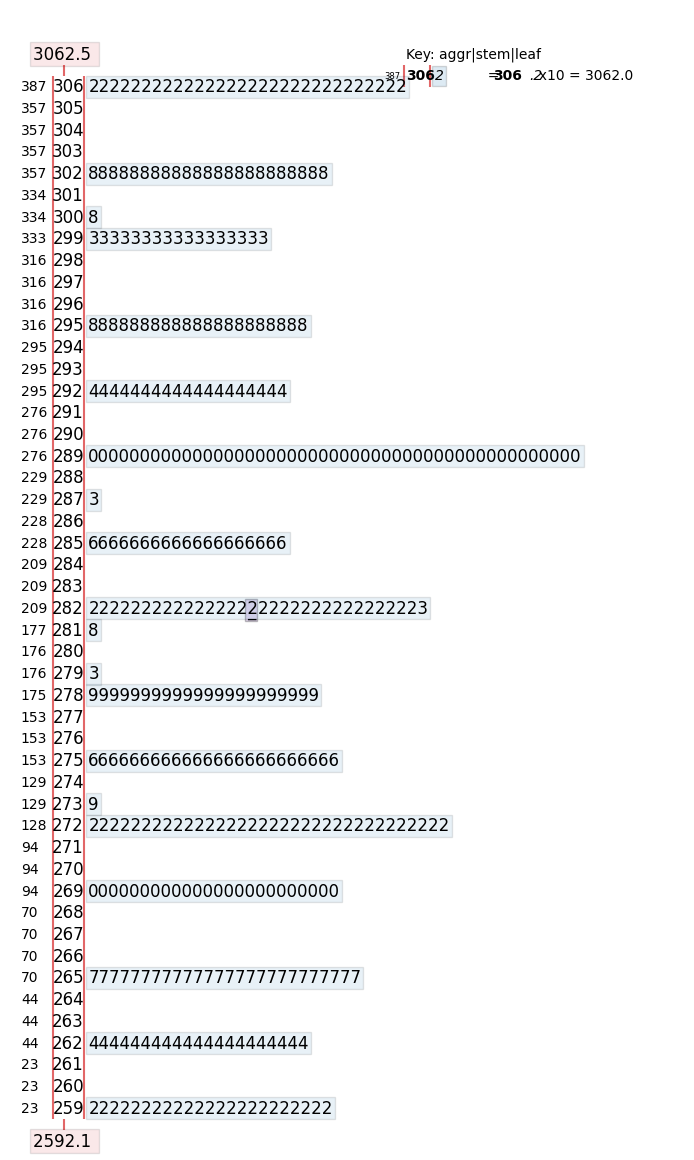
Range 5087.6-5587.6:

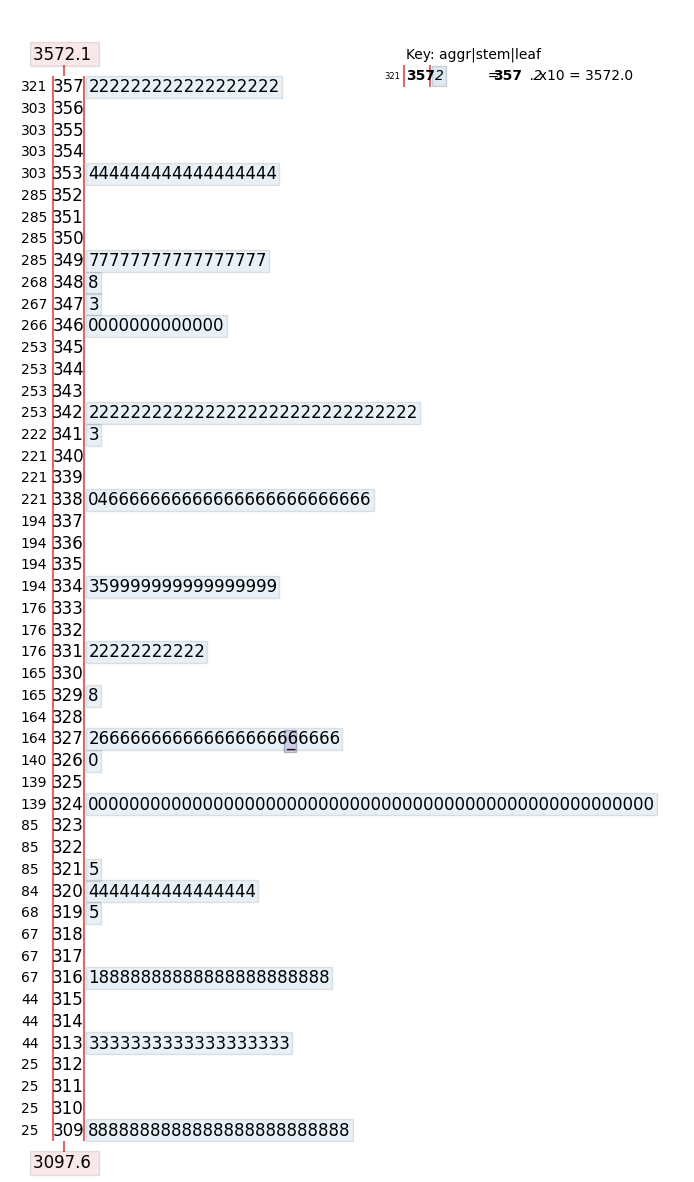
Range 5587.6-6087.6:

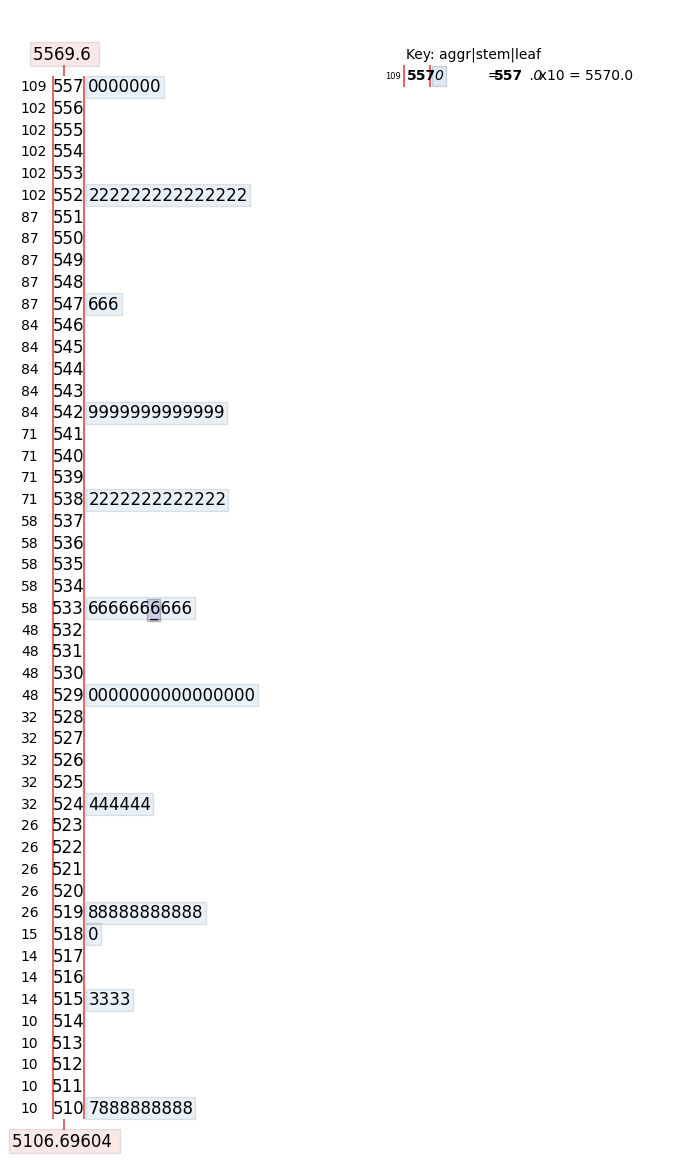
Median of entire dataset: 2560.0

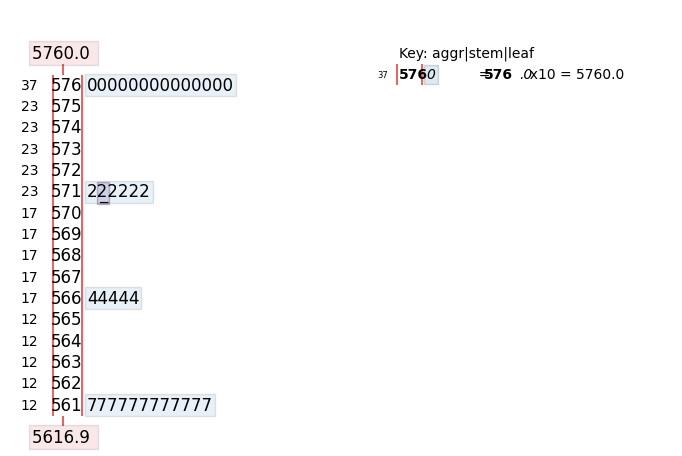












Interpretation

The median of the distribution is 2560.0 with around 10 to 15 different peak values at (2640,1390,3240,2592, 2251, 5402, 1280, 2332, 4000, 2720, 2892, ...) square meters, indicating various real estates developments in different regions in Melbourne. The greatest peak is at 1690 indicating an estate with same building area in square meters and having the maximum number of properties in Melbourne.

Dispersion

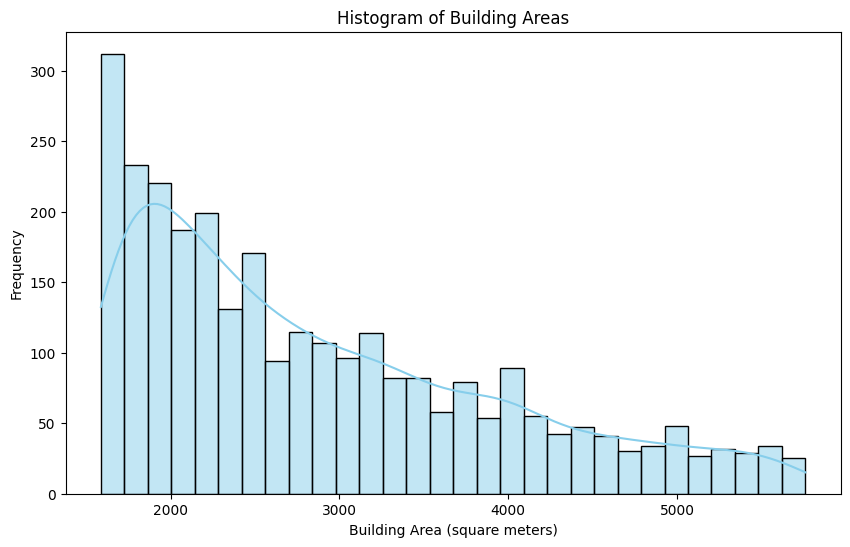
The range of the distribution is 4172.40 [square meters], indicating wide distribution of building areas. Majority of building areas are clustered around (1600-2600) range with a few moderate and few areas in the cluster around range (2700-6000).

Central Tendency

This clustering shows that, given a mean of 2875.6 [sq\_m], a mode of 1690, and median of 2560, it can be inferred that the majority of the building areas are below the mean, therefore real estate agents in Melbourne seem to have a common range of building area for their properties.

Symmetry

The distribution is not symmetric but right-skewed since most of the building areas are densely clustered around lower building areas and therefore it could inform us about the fact that most of the real estate properties in Melbourne are of relatively smaller building areas.

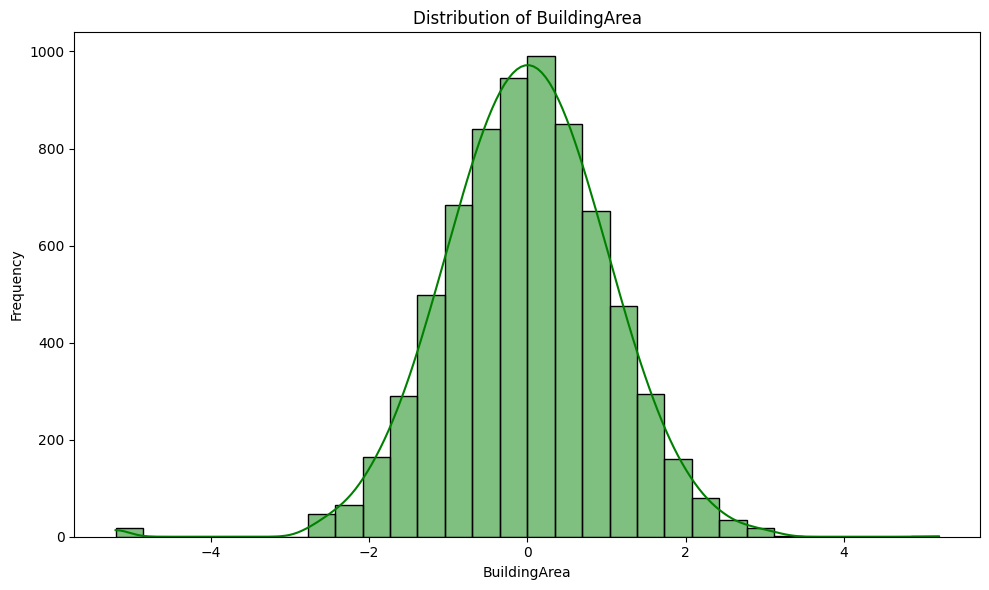


*Original distribution of variable*

The histogram of building areas presents a distribution where frequencies of different building sizes are plotted. Key observations:

* Peak Frequency: The highest frequency is for building areas around 1600-2000 square meters, with the highest bar at around 300 properties.
* Distribution Shape: The distribution shows a decreasing frequency as building areas increase, forming a right-skewed pattern rather than a normal distribution.
* Range: The building areas range from below 1000 square meters to over 5000 square meters.
* Frequency Decrease: The frequency of larger building areas steadily decreases, indicating fewer properties with larger sizes.
* Outliers: There are fewer properties with very large building areas (above 5000 square meters), showing a long tail on the right side.
* This right-skewed distribution suggests that while most properties have smaller building areas, there are some with significantly larger sizes making the distribution skewed to the right (positively-skewed).

#### (ii)



A Normal Distribution curve of the variable “Building Area” using Quantile Transformer.

The transformation is applied on each feature independently. First an estimate of the cumulative distribution function of a feature is used to map the original values to a uniform distribution. The obtained values are then mapped to the desired output distribution using the associated quantile function. Features values of new/unseen data that fall below or above the fitted range will be mapped to the bounds of the output distribution (*QuantileTransformer - Sklearn*, n.d.)

Interpretation

The histogram, has a bell-shaped curve, indicating that the distribution of building areas is approximately normal. This bell-shaped appearance is evidence of a normal distribution, indicating that the majority of building areas are clustered around a central value, with fewer observations as we move further from the center. In normal distribution, the mean, median, and mode tend to coincide, all residing at the center of the distribution. This central value represents the most common building area size, suggesting that typical buildings in Melbourne fall within the central range. The clustering of values around this central point indicates that most properties have building areas close to the mean, with deviations becoming less frequent as one moves away from the center. The distribution is not skewed to the left or right, showing that the building areas are normally distributed. The width of the histogram bars illustrates the variability or spread of the building areas. The data is symmetrically spread around the mean, indicating a consistent pattern of dispersion. One notable feature in the histogram is the presence of a small bar to the far left, which indicates some outliers with unusually low building areas. These outliers represent buildings that are significantly smaller than the majority of properties. While these values are much less common compared to the rest of the data points, their presence is important as it highlights the existence of a few properties that deviate markedly from the norm.

Tool: Python

Libraries:

import numpy as np

import pandas as pd

import plotly.express as px

from sklearn.preprocessing import QuantileTransformer

import plotly.graph\_objects as go

import matplotlib.pyplot as plt

import seaborn as sns

References

*An Introduction to Scikit-Learn: Machine Learning in Python*. (n.d.). Simplilearn.Com. Retrieved August 1, 2024, from https://www.simplilearn.com/tutorials/python-tutorial/scikit-learn

Chugani, V. (2024, March 6). Skewness Be Gone: Transformative Tricks for Data Scientists. *MachineLearningMastery.Com*. https://machinelearningmastery.com/skewness-be-gone-transformative-tricks-for-data-scientists/

*Melbourne Housing Snapshot*. (n.d.). Retrieved August 1, 2024, from https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot

*QuantileTransformer—Sklearn*. (n.d.). Retrieved August 1, 2024, from https://sklearn.vercel.app/docs/classes/QuantileTransformer

*Scikit-learn: Machine learning in Python—Scikit-learn 1.5.1 documentation*. (n.d.). Retrieved August 1, 2024, from https://scikit-learn.org/stable/

Shukla, P. (2022, October 28). Handling Missing Data with SimpleImputer. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2022/10/handling-missing-data-with-simpleimputer/