BUSINESS ANALYTICS REPORT

Milan Mitrovic | s4663796

Contents

Introduction

In the fiercely competitive realm of real estate consulting, securing a distinct advantage is paramount for success, the proposed solution: harnessing the power of business analytics & data mining. Employed as a business analyst in a real estate consulting firm with the goal of maximizing the firm's organisational performance with the provided Melbourne housing dataset to be used for leverage with business analytics. The Melbourne housing dataset contains 22 features: (id, rooms, price, method, type, seller, date, distance, region name, property count, bedroom, bathroom, car, land size, building area, council area, suburb, address, year built, latitude, longitude). The use of business analytics can significantly enhance our ability to extract valuable insights from the dataset provided. By conducting advanced analytical techniques and tools, we can identify hidden patterns, trends, and correlations within the data, enabling the firm to make informed decisions and drive business benefits. The business analytics process can help our firm find opportunities in data, predict trends that forecast future opportunities, and aid in selecting a course of action that optimizes the firm's allocation of resources to maximize value and performance.

By analyzing and interpreting the features in the provided dataset using various business analytics techniques such as data mining and data analysis, the real estate consulting firm can:

1. **Identify Market Trends and Opportunities:**
   - By analysing features such as property prices, methods of sale, and region names over time, a real estate consulting firm can identify trends in the housing market, such as increasing demand in certain regions or shifts in buyer preferences.
   - Insights from data analysis can unveil opportunities for investment, such as property types with high potential for appreciation or emerging neighbourhoods.
   - Acknowledging market trends allows the firm to foresee shifts in supply and demand, identify niche markets, and seize emerging opportunities.

2. **Gauge Property Values and Investment Potential**:
   - Analysing features on property characteristics (type, number of rooms, bathrooms, land size, building area) and transactional details (price, date, methods of sale) facilitate for the valuation of properties and the comparison with similar properties on the market.
   - Analysis of historical sales data and market indicators aids in assessing investment potential of properties, considering factors such as rental yields, capital growth prospects, and market stability.
   - By computing property values and investment returns, the firm can make informed decisions about buying, selling, or holding properties to maximize returns and minimize risks.

3. **Evaluating Pricing and Marketing Strategies**:
   - Utilizing data analysis guides optimal pricing strategies by considering factors such as property attributes, competitive pricing, and market demand.
   - Data analysis of attributes such as property type, size (rooms, bedrooms, bathrooms), and location (region name, suburb, distance from CBD) helps identify target demographics. For example, properties with larger land sizes and more rooms may appeal to families, while units or townhouses in urban areas may target investors.

- Insights from the data inform targeted marketing strategies by understanding buyer behaviour and preferences, the firm can tailor marketing efforts, prioritize listings, and allocate resources effectively to maximize exposure and sales opportunities.

4. **Improve Operational Efficiency and Decision-Making Processes:**
    - Data-driven decision-making optimizes operational processes, enhances resource allocation, and boosts overall efficiency within the firm.
    - Data analysis of property features such as rooms, bathrooms, car spots, land size, and building area can provide critical insights for enhancing operational efficiency and decision-making processes. Timely access to accurate property data empowers decision-makers to assess property values, prioritize listings, and allocate resources effectively to achieve business objectives.
    - By utilizing business analytics tools and techniques, the firm secures a competitive edge, swiftly adjusts to shifts in the market, and substantially enhances both operational efficacy and decision-making protocols.

5. **Supply Data-Driven Recommendations to Clients**:
    - Insights from data analysis enables the firm to supply clients with personalized recommendations that are customized to their individual needs, preferences, and risk profiles.
    - Utilizing data analysis on property attributes such as rooms, bathrooms, car spots, and property type, the firm can offer personalized recommendations tailored to their clients specific preferences and requirements. For instance, clients seeking spacious properties can be directed towards listings with more rooms and bathrooms, while those prioritizing convenient spaces may be recommended units or townhouses.
    - Data-driven recommendations may encompass property selection, financing options, investment strategies, and transactional timing based on market insights. By offering data-driven counselling: the firm enhances overall client satisfaction, builds trust, and reinforces long-term relationships, positioning the firm as a trusted consultancy in the real estate market.

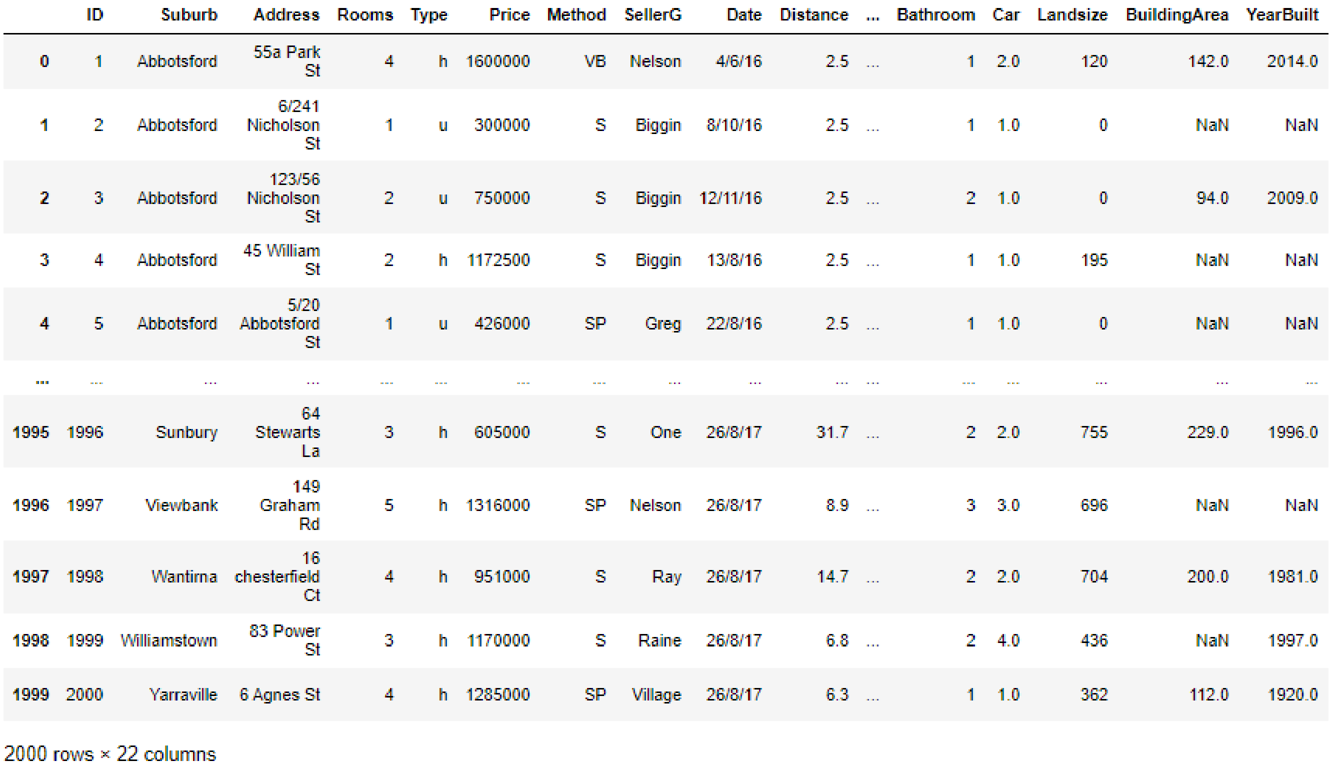
6. **Assess Risk Factors and Mitigate Potential Risks**:
    - Data analysis facilitates the identification and assessment of numerous risk factors that may impact property investments such as market volatility, economic shifts, and environmental risks.
    - By analyzing sales data features (price, method, date), the firm gains insights into market volatility, pricing trends, and transactional methods. This allows for the assessment of potential risks associated with price fluctuations and market shifts, enabling proactive risk management strategies.
    - Data analysis of property features (type, suburb, building area, distance from CBD, rooms, bathroom, car) helps identify properties at higher risk of depreciation or market saturation. This analysis informs investment decisions and risk mitigation measures by identifying properties with higher or lower risk levels based on their property characteristics.
    - By proactively identifying and addressing risks by developing risk mitigation strategies, the firm can safeguard investments, protect client interests, and maintain financial stability in the face of uncertainties.

Ultimately, business analytics can serve as an extremely useful and powerful tool for the real estate consulting firm: by harnessing the capabilities of business analytics, we can unlock the full potential of the Melbourne housing dataset provided, bringing substantial business benefits: Identification of market trends, evaluation of property values, optimization of pricing and marketing strategies, enhancement of operational efficiency and decision-making processes, provision of data-driven recommendations to clients, assessment of risk factors, supporting strategy development, improved portfolio management, and forecasting future business opportunities; subsequently maximizing organisational performance.

Overall, leveraging the Melbourne housing dataset for business analytics enables the real estate consulting firm to gain a competitive advantage - which is a necessity for success in the real estate market.

Data Cleaning/Task 1: Understand The Dataset

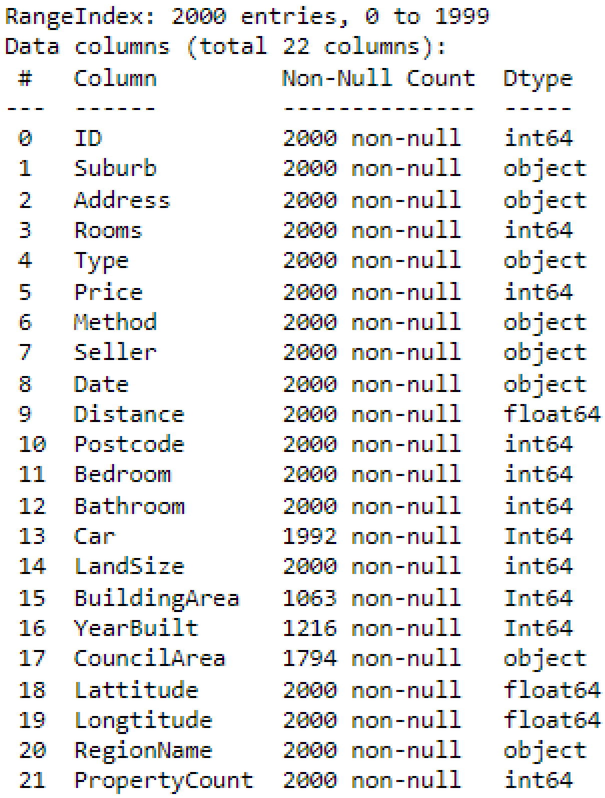
Before starting Exploratory Data Analysis (EDA) we need to prepare and clean the data, as dirty data can lead to misleading results and inaccurate conclusions.



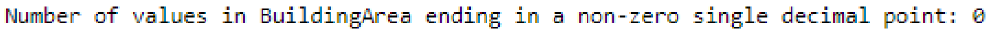
	ID	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	...	Bathroom	Car	Landsize	BuildingArea	YearBuilt
0	1	Abbotsford	55a Park St	4	h	1600000	VB	Nelson	4/6/16	2.5	...	1	2.0	120	142.0	2014.0
1	2	Abbotsford	6/241 Nicholson St	1	u	300000	S	Biggin	8/10/16	2.5	...	1	1.0	0	NaN	NaN
2	3	Abbotsford	123/56 Nicholson St	2	u	750000	S	Biggin	12/11/16	2.5	...	2	1.0	0	94.0	2009.0
3	4	Abbotsford	45 William St	2	h	1172500	S	Biggin	13/8/16	2.5	...	1	1.0	195	NaN	NaN
4	5	Abbotsford	5/20 Abbotsford St	1	u	426000	SP	Greg	22/8/16	2.5	...	1	1.0	0	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1995	1996	Sunbury	64 Stewarts La	3	h	605000	S	One	26/8/17	31.7	...	2	2.0	755	229.0	1996.0
1996	1997	Viewbank	149 Graham Rd	5	h	1316000	SP	Nelson	26/8/17	8.9	...	3	3.0	696	NaN	NaN
1997	1998	Wantirna	16 chesterfield Ct	4	h	951000	S	Ray	26/8/17	14.7	...	2	2.0	704	200.0	1981.0
1998	1999	Williamstown	83 Power St	3	h	1170000	S	Raine	26/8/17	6.8	...	2	4.0	436	NaN	1997.0
1999	2000	Yarraville	6 Agnes St	4	h	1285000	SP	Village	26/8/17	6.3	...	1	1.0	362	112.0	1920.0

2000 rows × 22 columns

Checking the dimensions of the dataset (2000 rows, 22 columns), as well as removing duplicates in case they were present (there weren't any).



```
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 22 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   ID             2000 non-null   int64  
 1   Suburb         2000 non-null   object 
 2   Address        2000 non-null   object 
 3   Rooms          2000 non-null   int64  
 4   Type           2000 non-null   object 
 5   Price          2000 non-null   int64  
 6   Method         2000 non-null   object 
 7   Seller         2000 non-null   object 
 8   Date           2000 non-null   object 
 9   Distance       2000 non-null   float64
 10  Postcode       2000 non-null   int64  
 11  Bedroom        2000 non-null   int64  
 12  Bathroom       2000 non-null   int64  
 13  Car            1992 non-null   Int64  
 14  LandSize       2000 non-null   int64  
 15  BuildingArea   1063 non-null   Int64  
 16  YearBuilt      1216 non-null   Int64  
 17  CouncilArea    1794 non-null   object 
 18  Lattitude      2000 non-null   float64
 19  Longtitude     2000 non-null   float64
 20  RegionName     2000 non-null   object 
 21  PropertyCount  2000 non-null   int64  
```



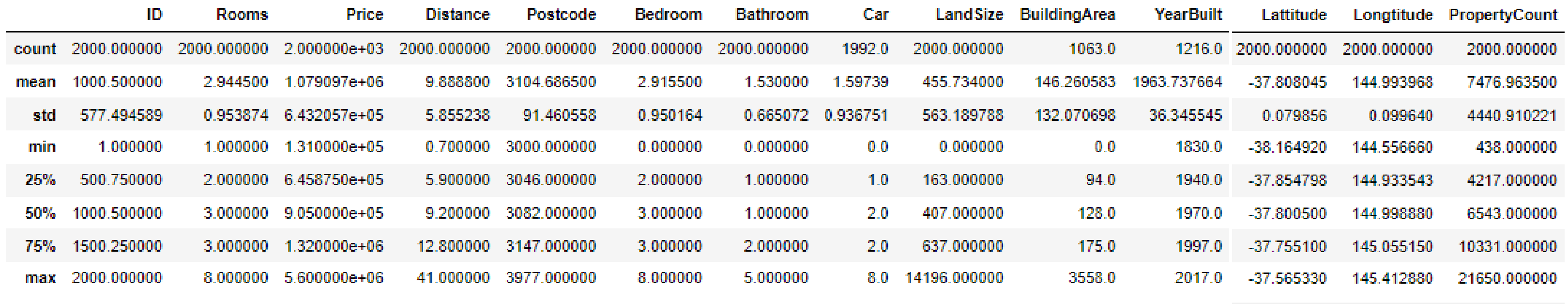
```
Number of values in BuildingArea ending in a non-zero single decimal point: 0
```

Changing the names of certain columns (SellerG to Seller, Bedroom2 to Bedroom, Landsize to LandSize, Regionname to RegionName, and Propertycount to PropertyCount) to follow the proper naming convention of this dataset (pascal case). Also, changing the data types of certain features to its appropriate data types: Car from float to integer, BuildingArea from float to integer, and YearBuilt from float to integer. We changed the data type of YearBuilt to its suitable data type integer as years are whole numbers and should not be represented as floating-point numbers, increasing conciseness, and ensuring consistency. For BuildingArea, this feature represents the building size in meters, so its original data type of float was fine for this however, there was actually no number of values within BuildingArea that ended in a non-zero decimal point, so we decided to change it to integer to improve its clarity. Finally, the Car feature represents the count of parking spots, it should ideally be an integer data type since you cannot have a fraction of a parking spot.

Since we have assigned the appropriate data types to our features, we can now identify the numeric and nominal features respectively:

- **Numerical**: ID, Rooms, Price, Distance, Postcode, Bedroom, Bathroom, Car, LandSize, BuildingArea, YearBuilt, Latitude, Longitude, PropertyCount.
- **Nominal**: Suburb, Address, Type, Method, Seller, Date, CouncilArea, RegionName.

There also appears to be null values in Car, BuildingArea, YearBuilt, and CouncilArea. We will deal with these null values after providing a simple statistical summary of our features.

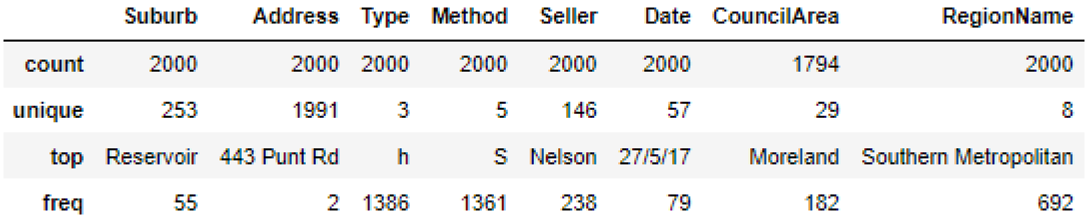


	ID	Rooms	Price	Distance	Postcode	Bedroom	Bathroom	Car	LandSize	BuildingArea	YearBuilt	Lattitude	Longtitude	PropertyCount
count	2000.000000	2000.000000	2.000000e+03	2000.000000	2000.000000	2000.000000	2000.000000	1992.0	2000.000000	1063.0	1216.0	2000.000000	2000.000000	2000.000000
mean	1000.500000	2.944500	1.079097e+06	9.888800	3104.686500	2.915500	1.530000	1.59739	455.734000	146.260583	1963.737664	-37.808045	144.993968	7476.963500
std	577.494589	0.953874	6.432057e+05	5.855238	91.460558	0.950164	0.665072	0.936751	563.189788	132.070698	36.345545	0.079856	0.099640	4440.910221
min	1.000000	1.000000	1.310000e+05	0.700000	3000.000000	0.000000	0.000000	0.0	0.000000	0.0	1830.0	-38.164920	144.556660	438.000000
25%	500.750000	2.000000	6.458750e+05	5.900000	3046.000000	2.000000	1.000000	1.0	163.000000	94.0	1940.0	-37.854798	144.933543	4217.000000
50%	1000.500000	3.000000	9.050000e+05	9.200000	3082.000000	3.000000	1.000000	2.0	407.000000	128.0	1970.0	-37.800500	144.998880	6543.000000
75%	1500.250000	3.000000	1.320000e+06	12.800000	3147.000000	3.000000	2.000000	2.0	637.000000	175.0	1997.0	-37.755100	145.055150	10331.000000
max	2000.000000	8.000000	5.600000e+06	41.000000	3977.000000	8.000000	5.000000	8.0	14196.000000	3558.0	2017.0	-37.565330	145.412880	21650.000000

The above table depicts a simple statistical summary of every numerical feature: displaying the count, mean, standard deviation, minimum value, (25%, 50% (or median), 75%) percentile values, and the maximum value of each and every numerical feature.

1. **Count**: The number of non-null values in each column.

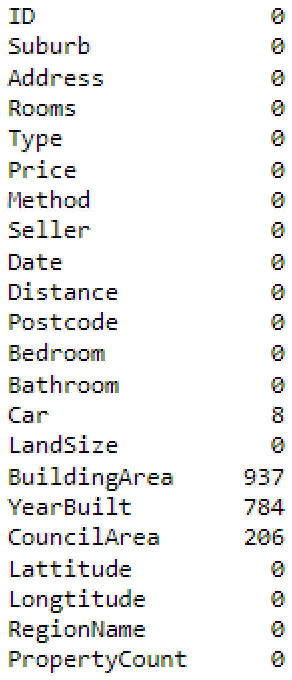
2. **Mean**: The average (mean) of each column.
3. **Standard Deviation (std)**: The measure of the amount of variation from their mean values in each column.
4. **Minimum (min)**: The smallest value in the column.
5. **25th Percentile (25%)**: The value below which 25% of the data fall.
6. **50th Percentile (50%) or Median**: The middle value of the dataset.
7. **75th Percentile (75%)**: The value below which 75% of the data fall.
8. **Maximum (max)**: The largest value in the column.



	Suburb	Address	Type	Method	Seller	Date	CouncilArea	RegionName
count	2000	2000	2000	2000	2000	2000	1794	2000
unique	253	1991	3	5	146	57	29	8
top	Reservoir	443 Punt Rd	h	S	Nelson	27/5/17	Moreland	Southern Metropolitan
freq	55	2	1386	1361	238	79	182	692

The above table highlights the simple statistical summary of every nominal feature: showing the count, unique, top, and the frequency of every single nominal feature.

1. **Count**: The number of non-null values in each column.
2. **Unique**: The number of unique values in each column.
3. **Top**: The most frequently occurring value in each column.
4. **Freq**: The frequency or number of the top value in each column.

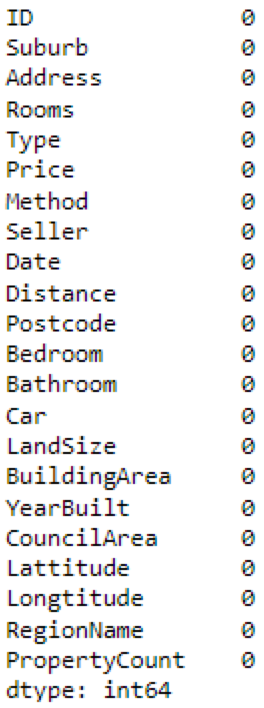


```
ID                 0
Suburb             0
Address            0
Rooms              0
Type               0
Price              0
Method             0
Seller             0
Date               0
Distance           0
Postcode           0
Bedroom            0
Bathroom           0
Car                8
LandSize           0
BuildingArea     937
YearBuilt        784
CouncilArea      206
Lattitude          0
Longtitude         0
RegionName         0
PropertyCount      0
```

8 null values in Car, 937 null values in BuildingArea, 784 null values in YearBuilt and 206 null values in CouncilArea. There are many missing values present. Significant missing values in BuildingArea, YearBuilt, and CouncilArea. Insignificant missing values in Car. I deal with the insignificant and significant null values as follows:

- Insignificant null values: I deal with the insignificant null values in Car by removing the 8 rows which contain null values. We decided to do this as it was only 8 rows removed in total, which equates to 0.4% of the entire dataset: insignificant, maintaining data integrity and quality - removing nulls.

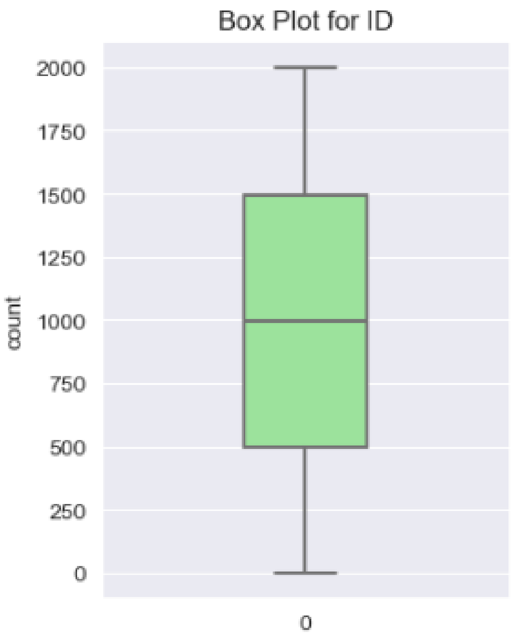
- Significant null values: I deal with the significant null values in BuildingArea by replacing them with the median values from its column. We do this to preserve the 937 rows or approximately 47% of the dataset - which is significant. We choose median here specifically because its robust to outliers and BuildingArea contains outliers; overall, this approach eliminates the null values, increasing data integrity and quality. For YearBuilt we replaced the null values

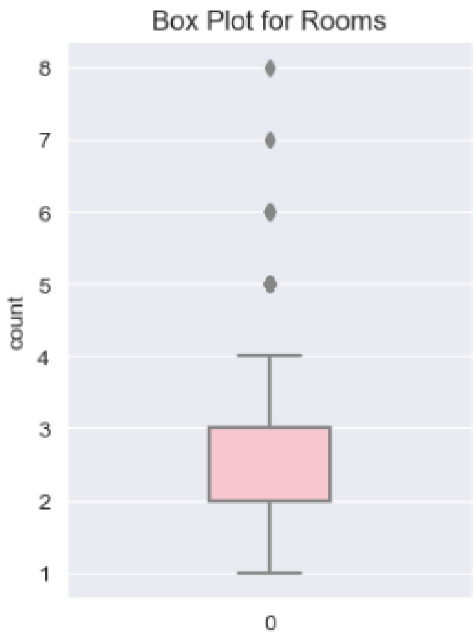


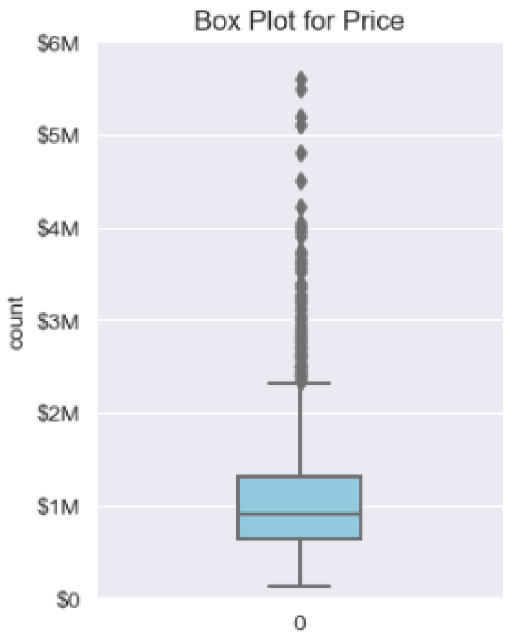
```
ID               0
Suburb           0
Address          0
Rooms            0
Type             0
Price            0
Method           0
Seller           0
Date             0
Distance         0
Postcode         0
Bedroom          0
Bathroom         0
Car              0
LandSize         0
BuildingArea     0
YearBuilt        0
CouncilArea      0
Lattitude        0
Longtitude       0
RegionName       0
PropertyCount    0
dtype: int64
```

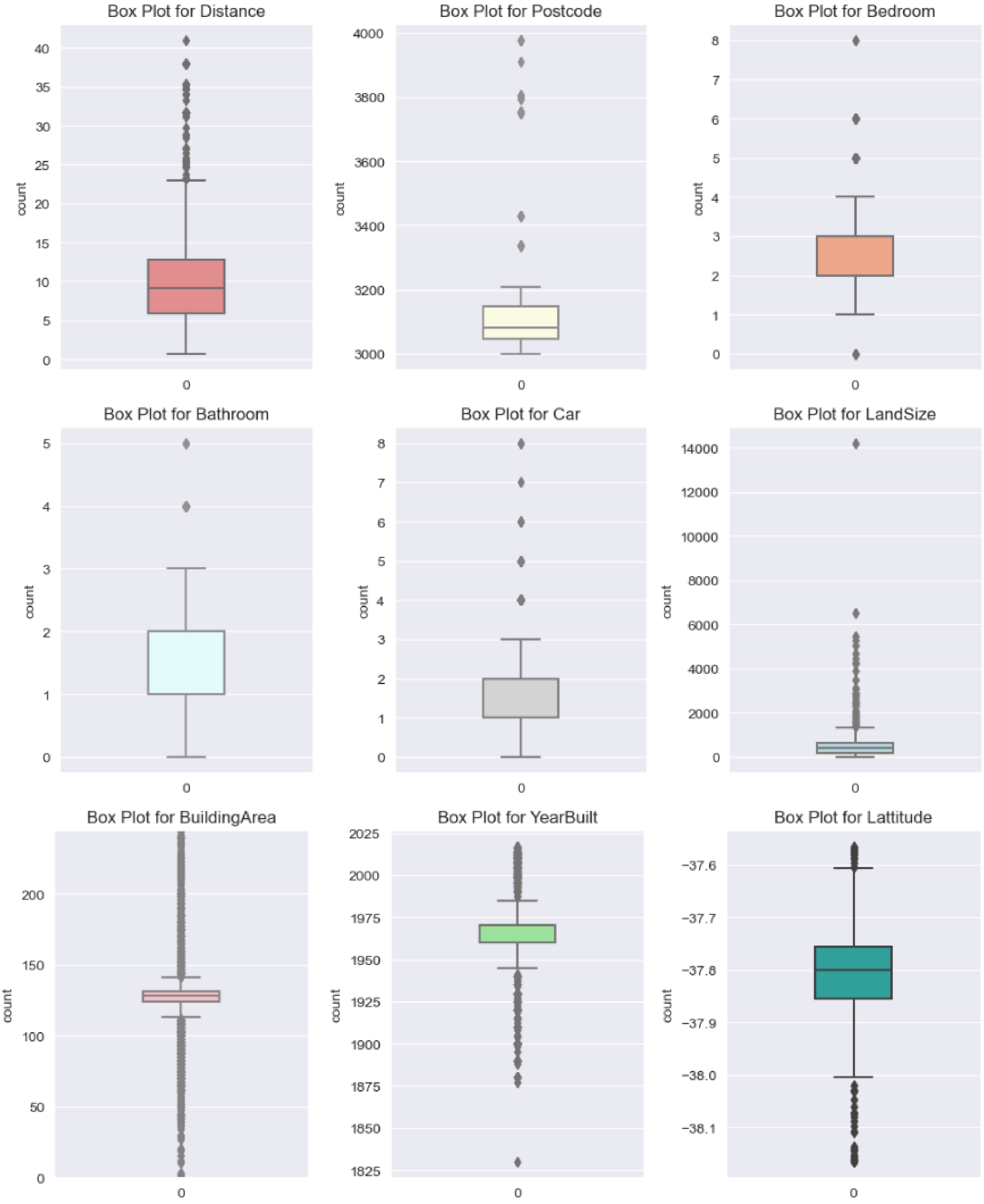
with the mode of its column. This is done to deal with the nulls, maintain data integrity and quality, as well as preserving 784 rows or around 39.4% of the dataset. Finally, for CouncilArea we also replaced the 206 null values with the mode of itself. Again, this is done to preserve the data frame (10.35% of the data frame) and enhance the data integrity and quality of the dataset by getting rid of the nulls. We opt for mode here because CouncilArea is a nominal attribute, and the mode is the most appropriate measure here, as mean or median is not applicable to nominal attributes. This process was a necessity as the null values would have distorted the data - negatively impacting the analysis.

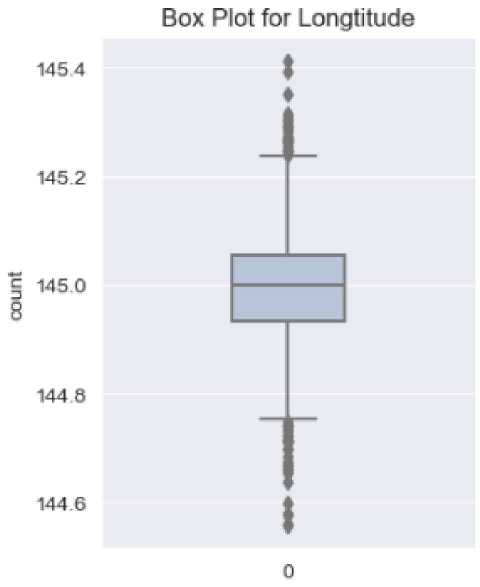
Now that we've addressed the null values in the dataset, it's time to investigate the outliers in all numerical columns. This will be achieved by data visualization techniques – specifically boxplots. Boxplots displays several key descriptive statistics, including the median, quartiles, and potential outliers, in a concise and easy-to-understand manner. However, we will only be focusing on outliers here as this is a part of the data cleaning process.

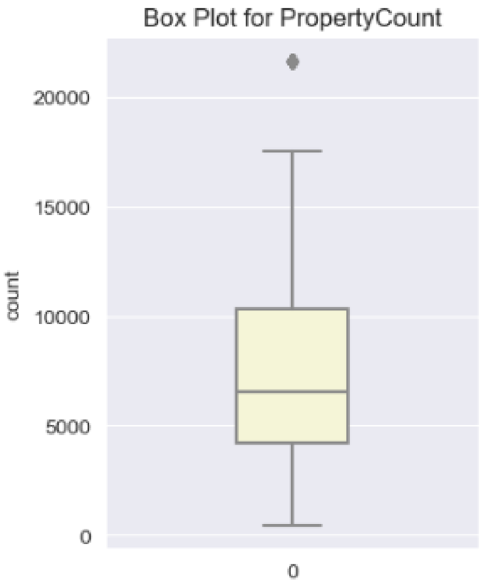







Box Plot for Distance
count
0
5
10
15
20
25
30
35
40
0
Box Plot for Postcode
count
3000
3200
3400
3600
3800
4000
0
Box Plot for Bedroom
count
0
1
2
3
4
5
6
7
8
0
Box Plot for Bathroom
count
0
1
2
3
4
5
0
Box Plot for Car
count
0
1
2
3
4
5
6
7
8
0
Box Plot for LandSize
count
0
2000
4000
6000
8000
10000
12000
14000
0
Box Plot for BuildingArea
count
0
50
100
150
200
0
Box Plot for YearBuilt
count
1825
1850
1875
1900
1925
1950
1975
2000
2025
0
Box Plot for Lattitude
count
−38.1
−38.0
−37.9
−37.8
−37.7
−37.6
0

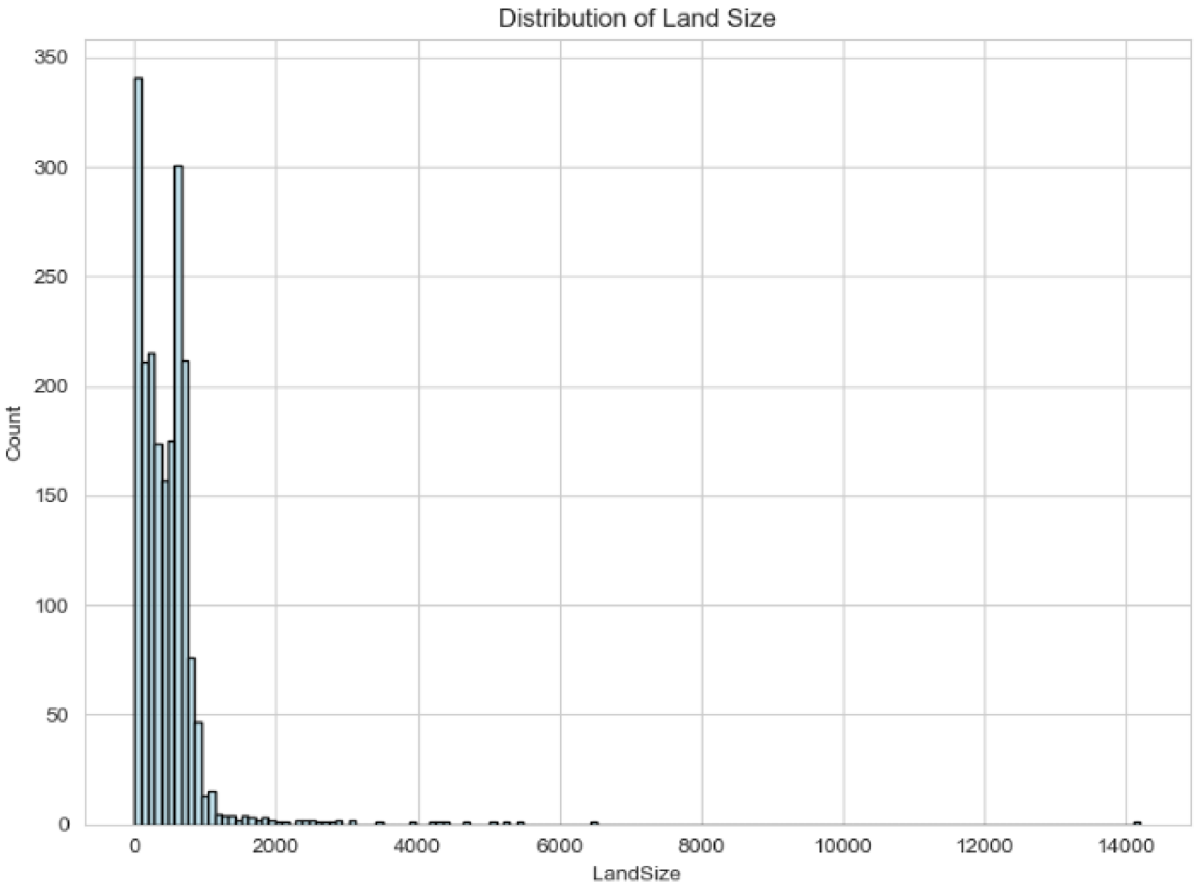


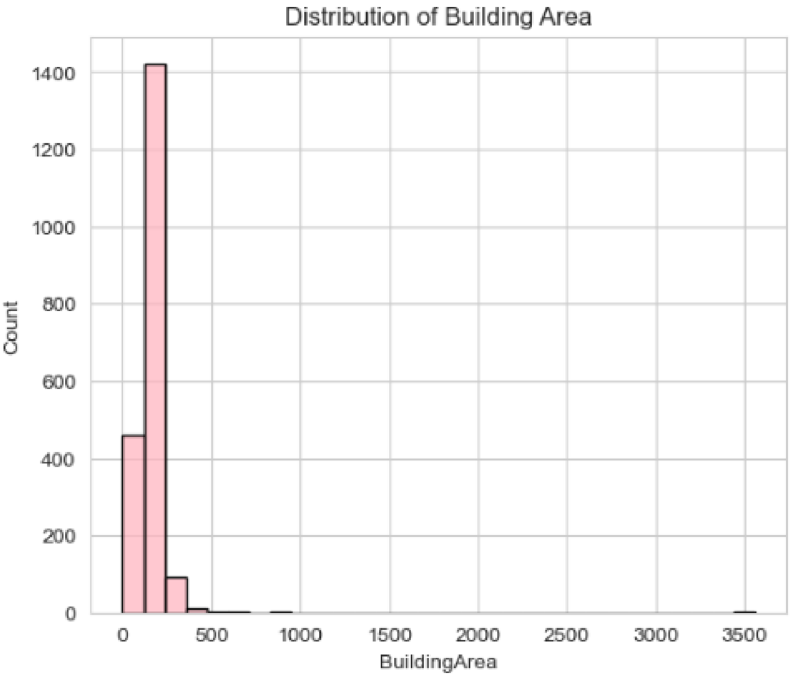


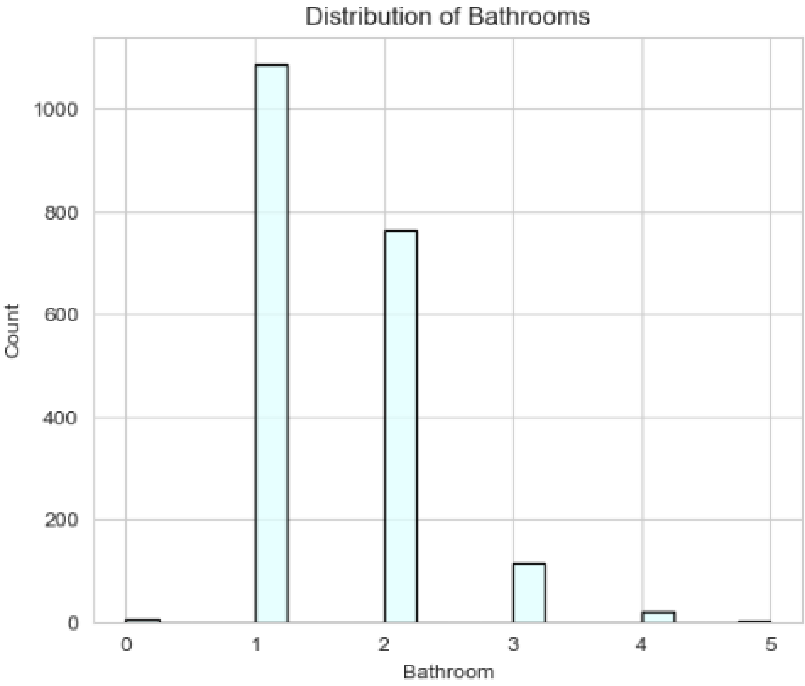
(Rooms, Bedroom, Bathroom, Car, YearBuilt boxplot medians aren't showing properly because their medians and Q1/Q3 scores are equal – Q1/Q3 is blocking the median).

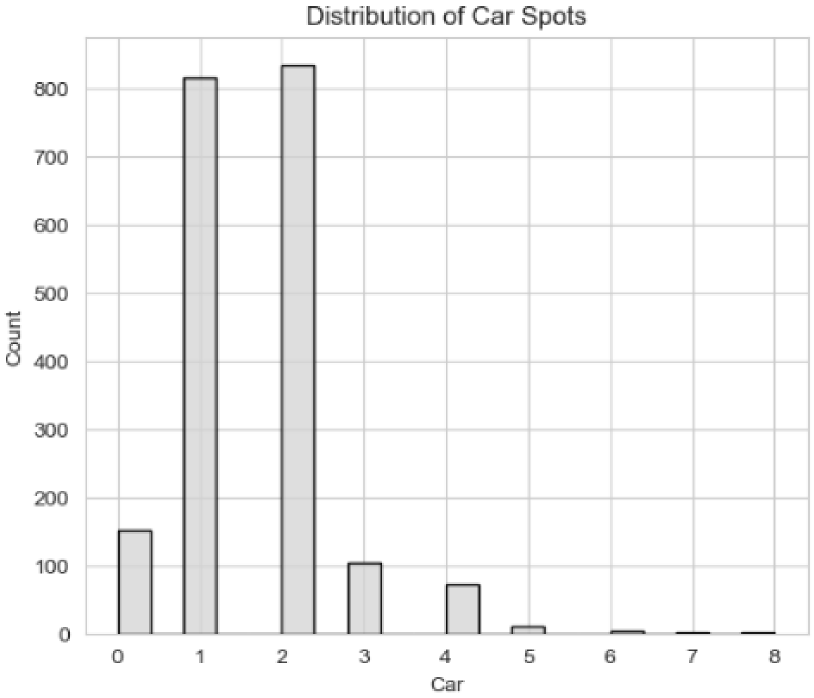
Clearly, apart from ID there are outliers present in every other attribute. The outliers in Rooms, Price, Distance, Postcode, YearBuilt, Latitude, Longitude, and PropertyCount all appear to be valid extreme values as they are realistic. Bedroom, Bathroom, and Car upper outliers also appear to be valid extreme values as 8 bedrooms, 5 bathrooms, and 8 car spots are feasible. We can conclude that the outliers covered so far are valid values that aren't errors. However, the extreme upper outliers for LandSize and BuildingArea are questionable, at a land size of 14196 square meters and a building size of 3558 square meters (which is not visible in the BuildingArea boxplot because the disparity between the max value/outlier and the other corresponding values of the boxplot are so vast the boxplot ceases to display correctly if we were to plot y ticks to 3600). The LandSize extreme upper outlier (14196m^2) is actually a valid extreme value – we verified the addresses land size (52/73 River St, Richmond). However, the BuildingArea extreme upper outlier (3558m^2) is a data entry error we checked the address (186 Queens Pde, Fitzroy North) and validated the building size at 145m^2 – we will deal with this outlier once we have finalized the outlier/0-value validation process.

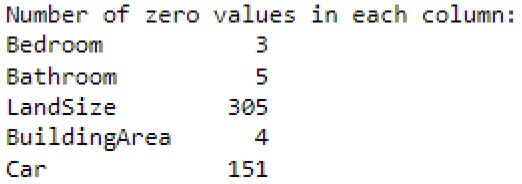
Bedroom and BuildingArea seem to contain 0-value outliers which doesn't make sense and needs more investigating. The 0-values for Bathroom, Car, and LandSize are not showing up because the boxplots lower whisker range starts at 0 for these features, so these 0-values are not being displayed as lower outliers by default because of this. This is a problem because if we refer to the simple statistical summary of numerical features table, we can see these columns contain 0-values as minimum values which are clearly errors. We will display these particular features (Bathroom, Car, LandSize) in histograms instead to avoid this, so we can better visualize the errors. As well as visualizing the highest upper outlier in BuildingArea which wasn't possible with the boxplot for BuildingArea.


Distribution of Land Size
Count
350
300
250
200
150
100
50
0
0
2000
4000
6000
8000
10000
12000
14000
LandSize


Distribution of Building Area
Count
1400
1200
1000
800
600
400
200
0
0
500
1000
1500
2000
2500
3000
3500
BuildingArea


Distribution of Bathrooms
Count
1000
800
600
400
200
0
0
1
2
3
4
5
Bathroom





```
Number of zero values in each column:
Bedroom            3
Bathroom           5
LandSize         305
BuildingArea       4
Car              151
```

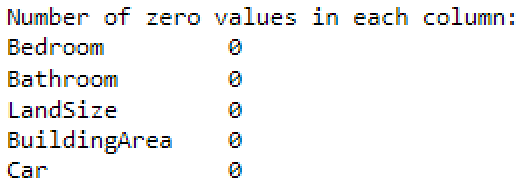
Now that we can visualize the extreme upper outlier in BuildingArea and 0-values in Bathrooms, Car, and LandSize we will begin with the final stages of the data cleaning process. By observing the above histograms or description of the zero values in each column: we can ascertain the number of 0-values in each column. Insignificant amounts in Bedroom (3), Bathroom (5), and BuildingArea (4). Significant amounts in LandSize (305) and Car (151).

We deal with the described issues above accordingly:

- **Insignificant outlier values/0-values:** We begin by removing the 3558m^2 extreme upper outlier we identified as an error from BuildingArea, along with its lower outlier 0-values, as well as the remaining 0-values in Bedroom, Bathroom. We do this because these 0-values are junk values which would skew the data; negatively impacting the data analysis. Essentially, we are removing noise as it is impossible for a house in urban Melbourne to have a 0 square meter building size, along with 0 bedrooms. The same way it is impossible for a house, unit, or townhouse to have 0 bathrooms - properties are expected to have at least 1 bathroom as bathrooms are considered essential features in properties, as they provide necessary facilities for personal hygiene and sanitation. Therefore, we can conclude these 0-values/outliers are data entry errors, and as they only total for 13 rows (less than 1% of the dataset) we go with the practical way of removing them.

- **Significant 0-values:** We decided to deal with the significant 0-values in LandSize and Car by replacing them with their own median. This is mainly done to preserve the dataset: 456 rows in total - approximately 23% of the total dataset. However, we also deal with the junk 0-values at the same time. In urban areas like Melbourne, Australia, it's highly unlikely for properties such as houses, units, or townhouses to have zero car spots, as parking spots are an essential amenity. Therefore, these 0-values in the Car column are data entry errors. Again, the same applies to LandSize: land size is a fundamental attribute of properties, it's impossible for properties such as houses, units, or townhouses to have zero land size, as even the smallest properties would have some land size. - 0-values in the LandSize column are data entry errors.

Overall, by dealing with these 0-values, we improve the data quality and integrity of the dataset, in addition we rectify the data entry 0-value errors (noise) which would have otherwise skewed the data, adversely affecting the data analysis and the quality of the dataset.

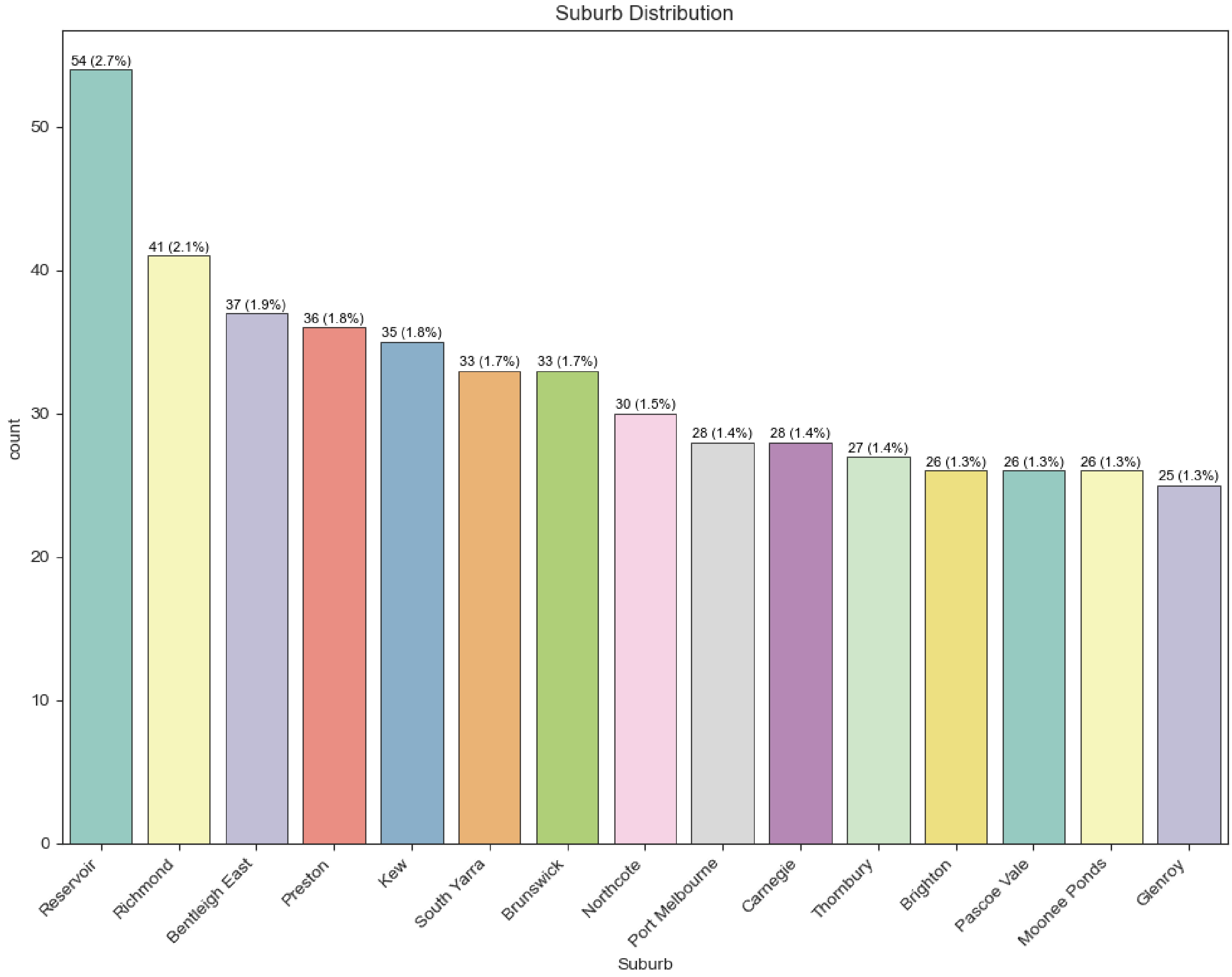


```
Number of zero values in each column:
Bedroom         0
Bathroom        0
LandSize        0
BuildingArea    0
Car             0
```

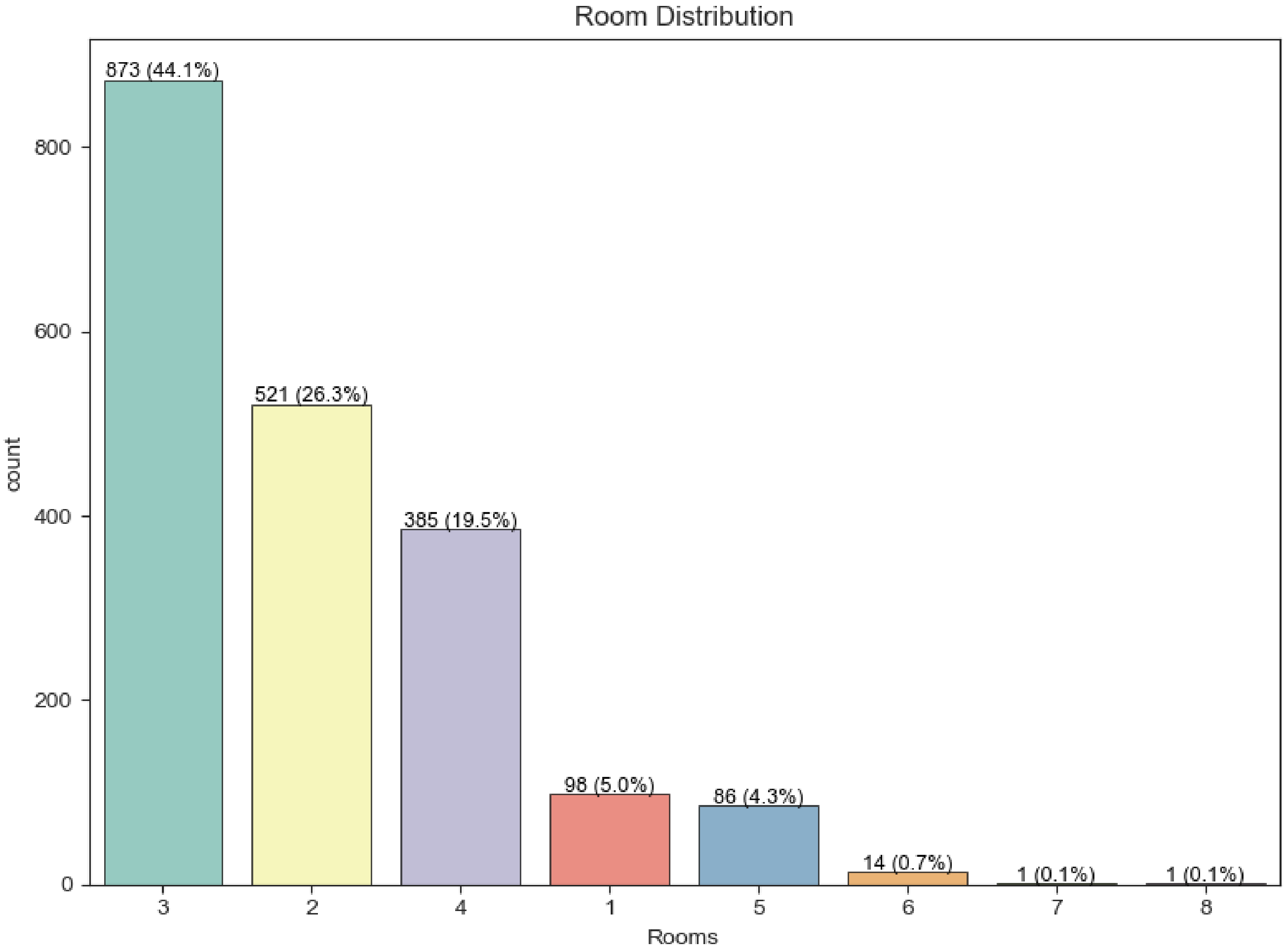
Updated number of zero values in each column, we can clearly observe the outliers/0-values which were errors are removed. Now that the data cleaning process is over, we can now move on to Exploratory Data Analysis (EDA).

Exploratory Data Analysis

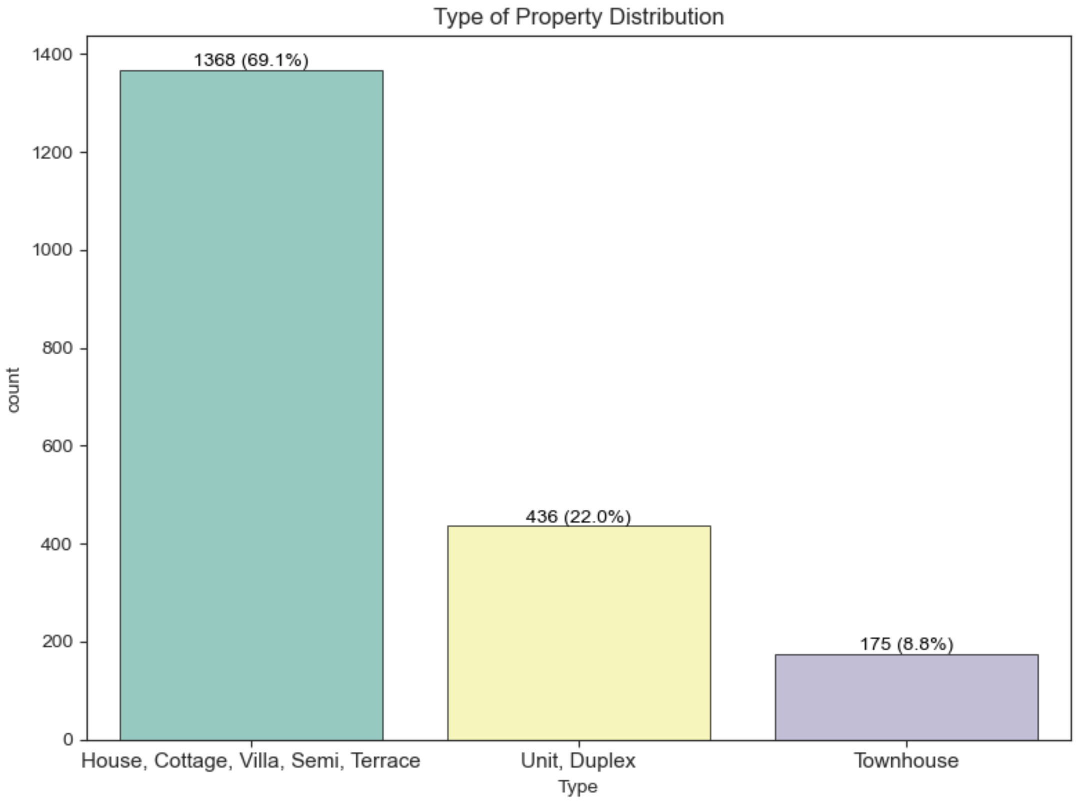
Exploratory Data Analysis (EDA): the process of analyzation and visualization to summarize the attributes and their corresponding relationships, to gain valuable insights about the dataset.



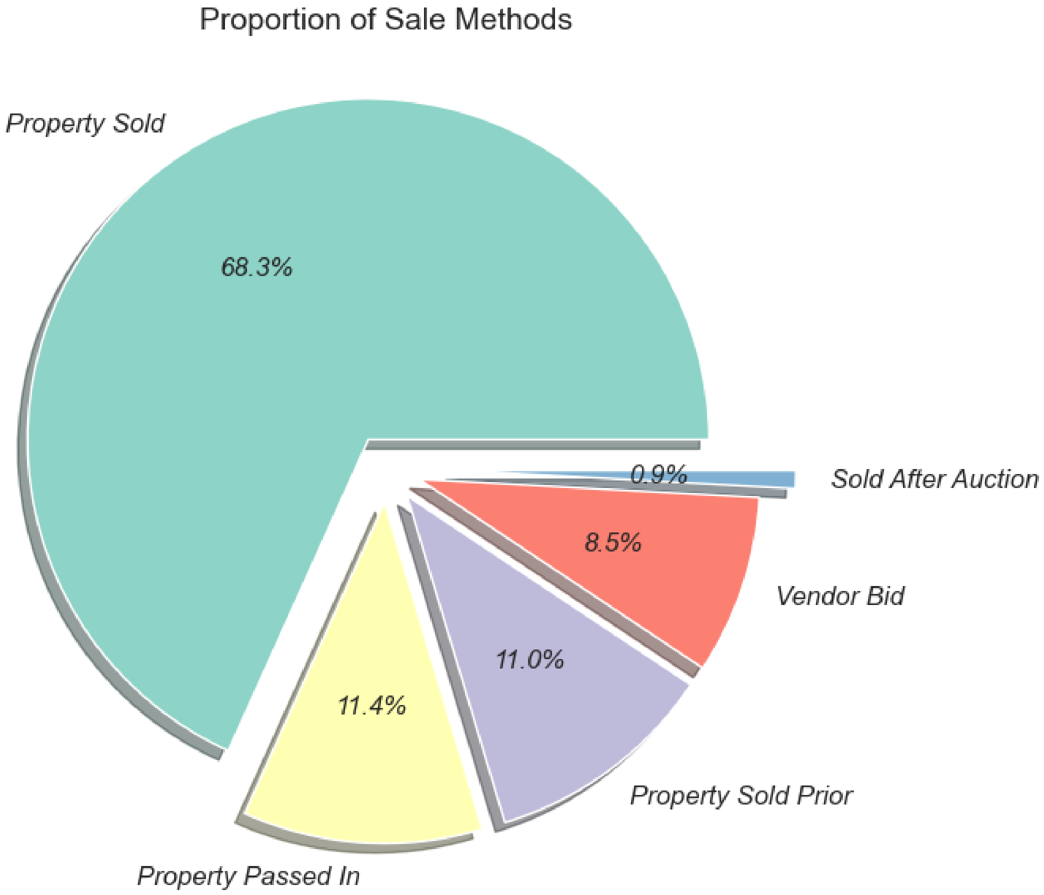
This count plot displays the count and the proportion of the Suburb feature, it depicts the top 15 suburbs in descending order within the dataset. The most prevalent suburb is Reservoir at a count of 54 or proportionally speaking 2.7% of the data frame. In contrast, Glenroy is last at a count of 25 (1.3% in proportion).



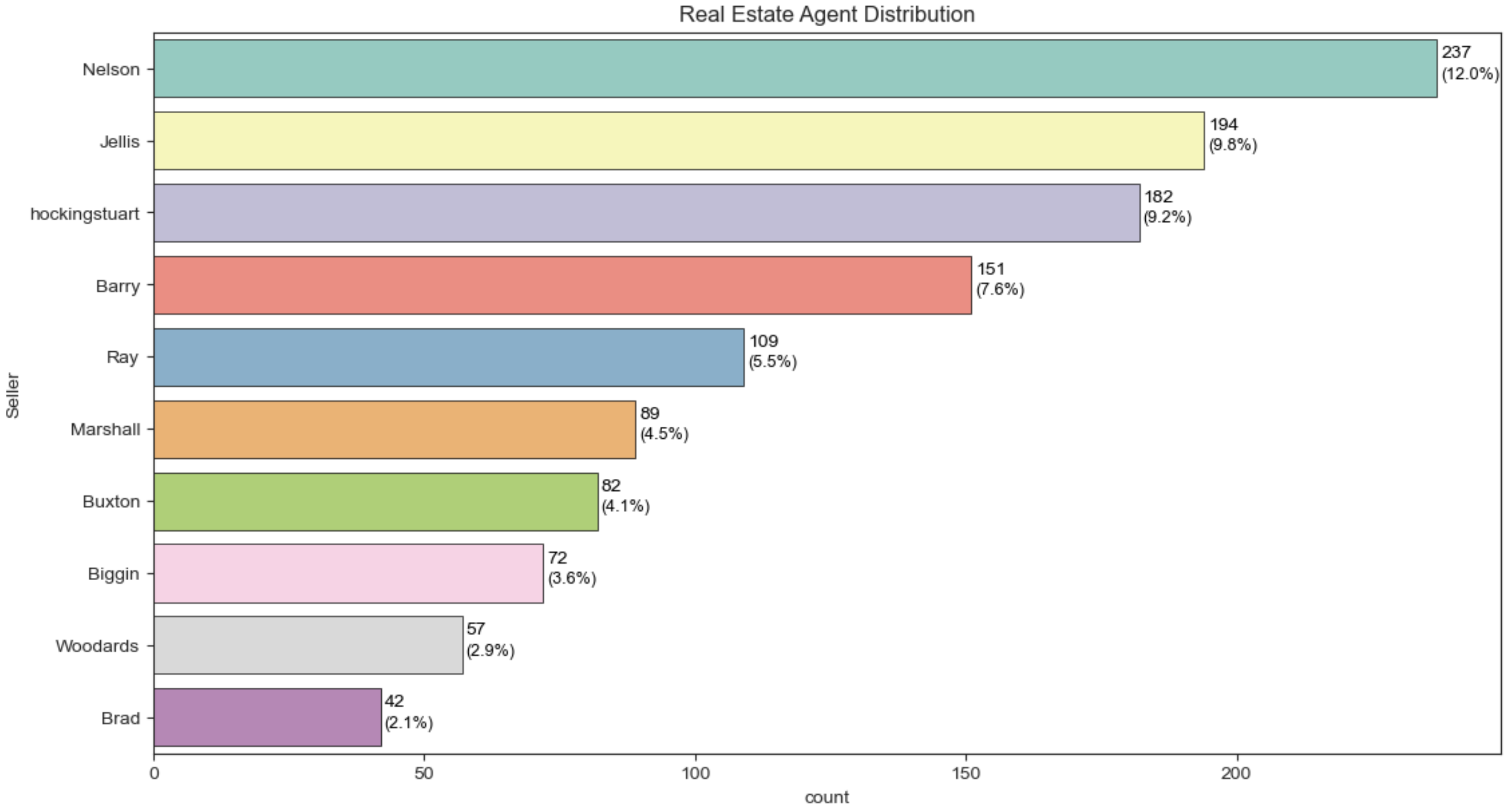
This count plot depicts the count and the proportion of Rooms. It highlights the most common number of rooms within properties in this dataset in descending order. The majority of the properties within the dataset have 3 rooms: at a count of 873 or 44.1% in proportional terms. The complete opposite can be said about properties with 8 rooms: amounting to only a count of 1 or 0.1% of the dataset.



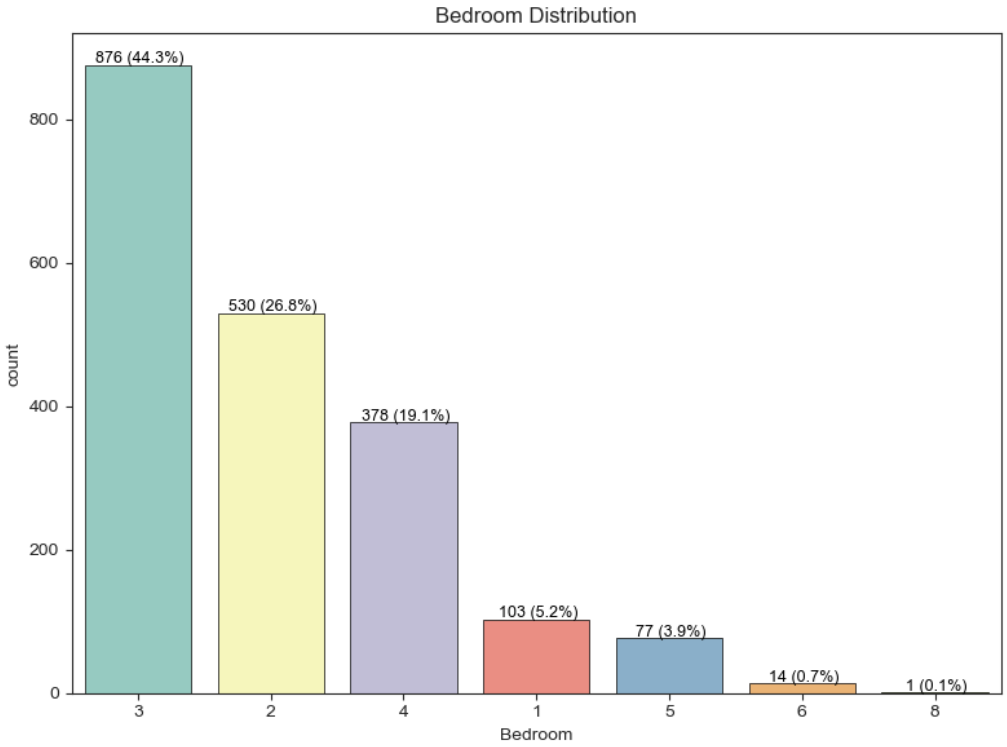
This count plot highlights the count and proportion of the type of properties in the dataset. It is displayed in descending order. Houses, cottages, villas, semi-detached houses, and terraces are the most common type of properties, at a count of 1368 or 69.1% of the dataset. The runner-up being units and duplexes, having a count of 436 or 22% in proportional terms. The least amount of properties within the data frame are townhouses: 175 in total or 8.8%.



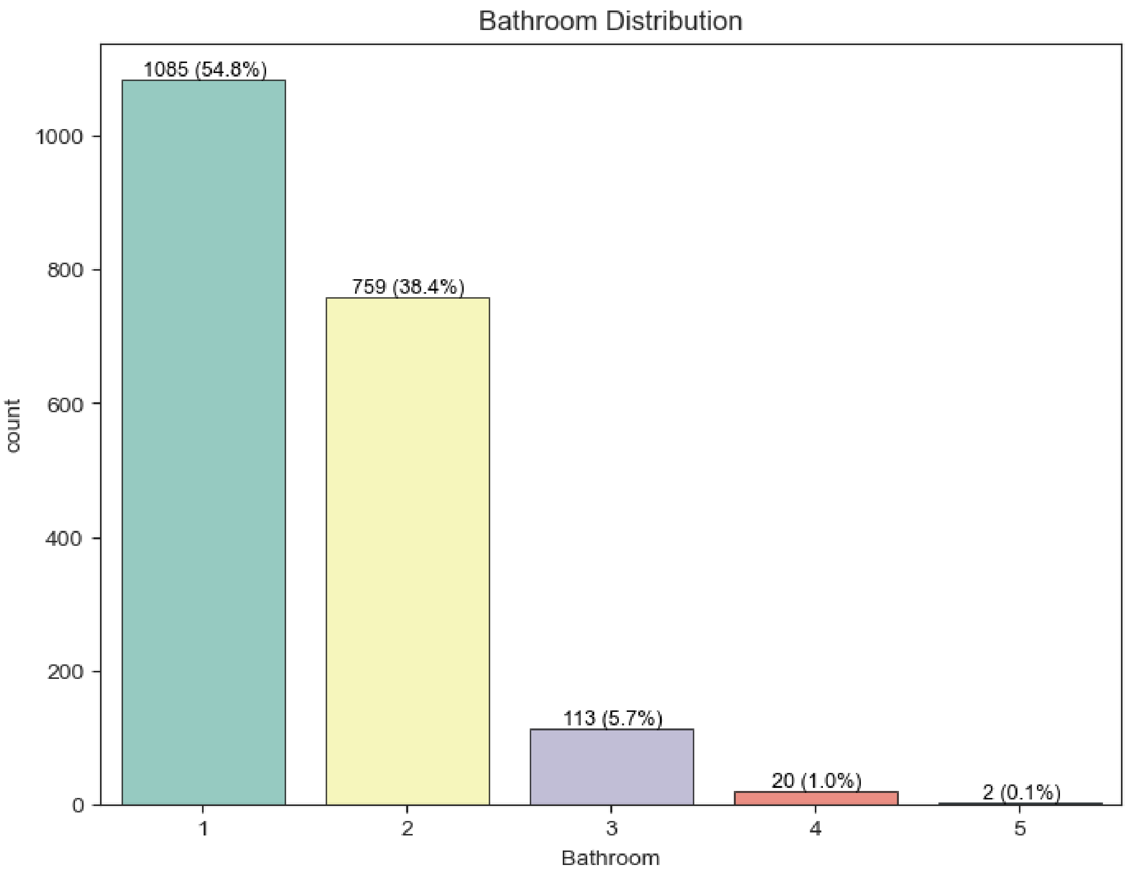
This pie plot displays the proportion of the Method feature. Essentially, it displays the most common types of method sale types within the dataset. Property Sold has the highest proportion by far: 68.3%. In contrast, Sold After Auction has the lowest proportion at 0.9% respectively.



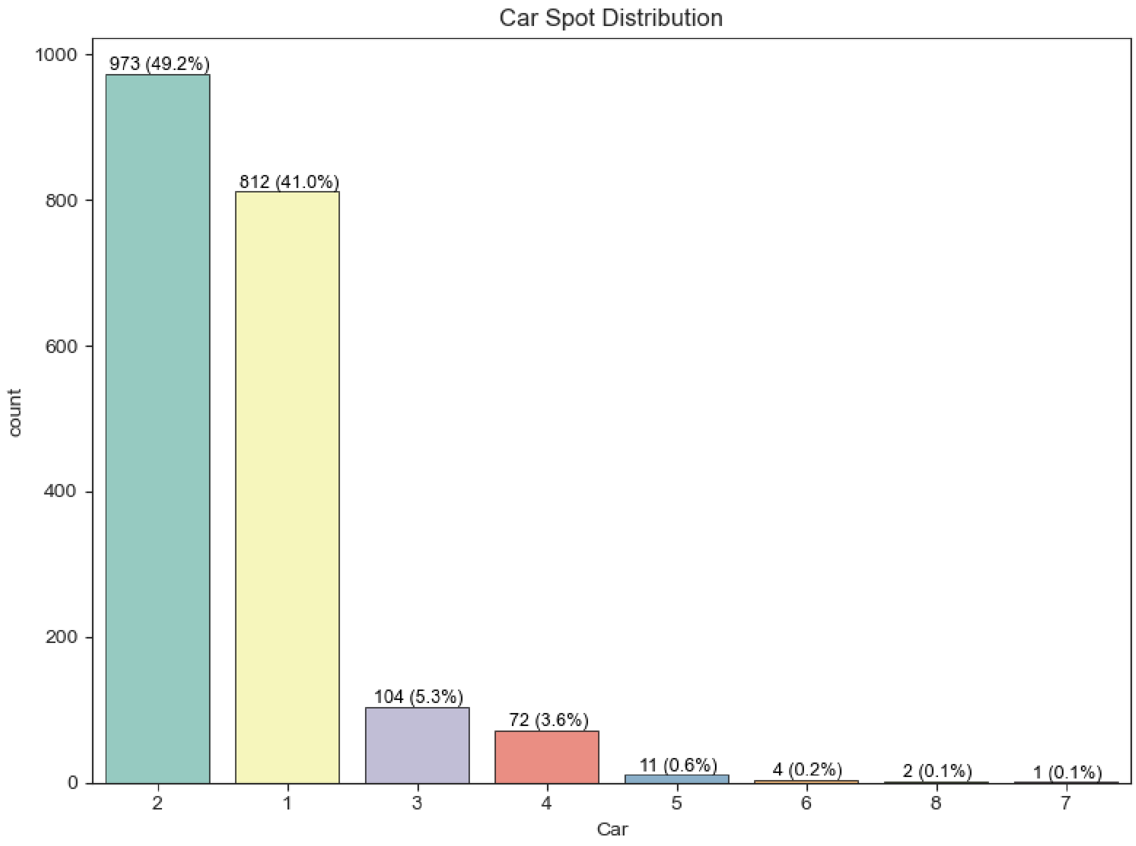
The above count plot represents the distribution and proportion of real estate agents; specifically, the top 10 real estate agents in descending order within the dataset. Nelson is the best real estate agent, having a count of 237 and the proportion of 12% within the dataset. Conversely, Brad is the worst real estate agent with a count of 42 or a proportion of 2.1%.



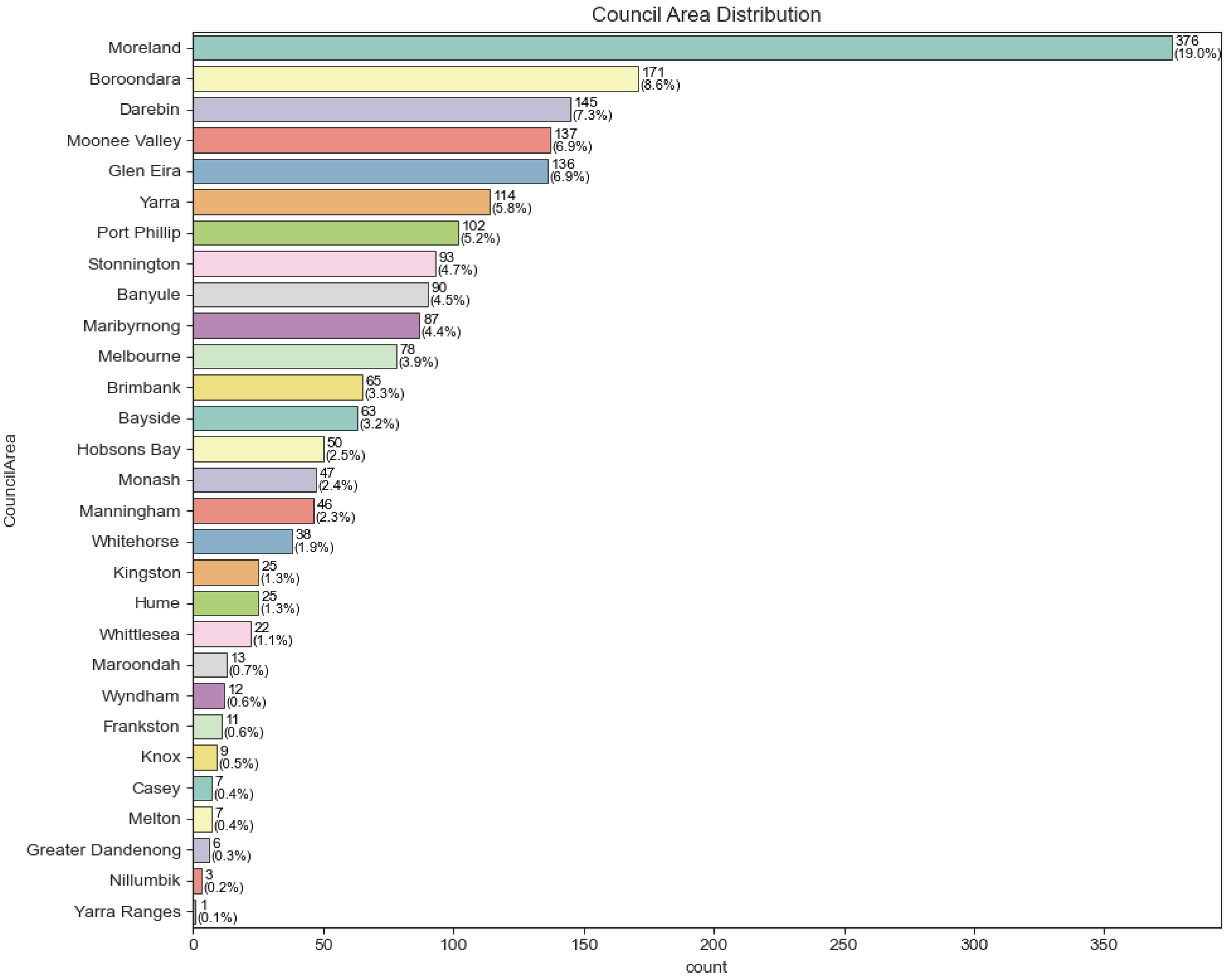
This count plot illustrates the distribution of Bedroom. Basically, it shows the majority numeric value of bedrooms within properties in descending order inside the dataset. A property that has 3 bedrooms is the most common, with a count of 876 or 44.3% of the dataset. In comparison, a property that has 8 bedrooms is the least common at just 1 count or 0.1% proportionally speaking.



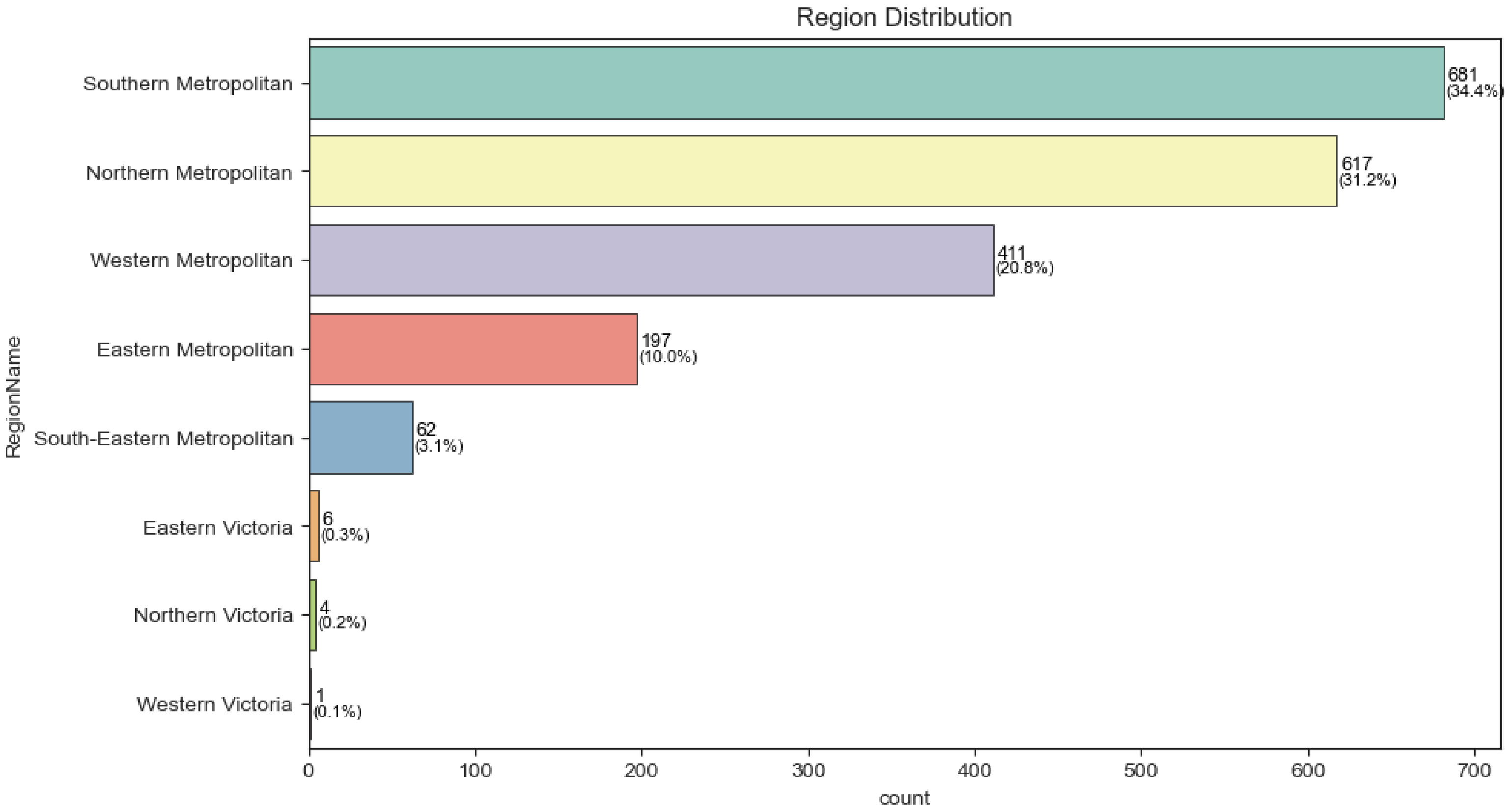
This count plot depicts the distribution of Bathroom within properties in descending order within the dataset. Properties with 1 bathroom are the most prevalent, containing a count of 1085 or 54.8% proportionally. In contrast, properties with 5 bathrooms are the smallest, at a count of 2 or 0.1% of the data frame.



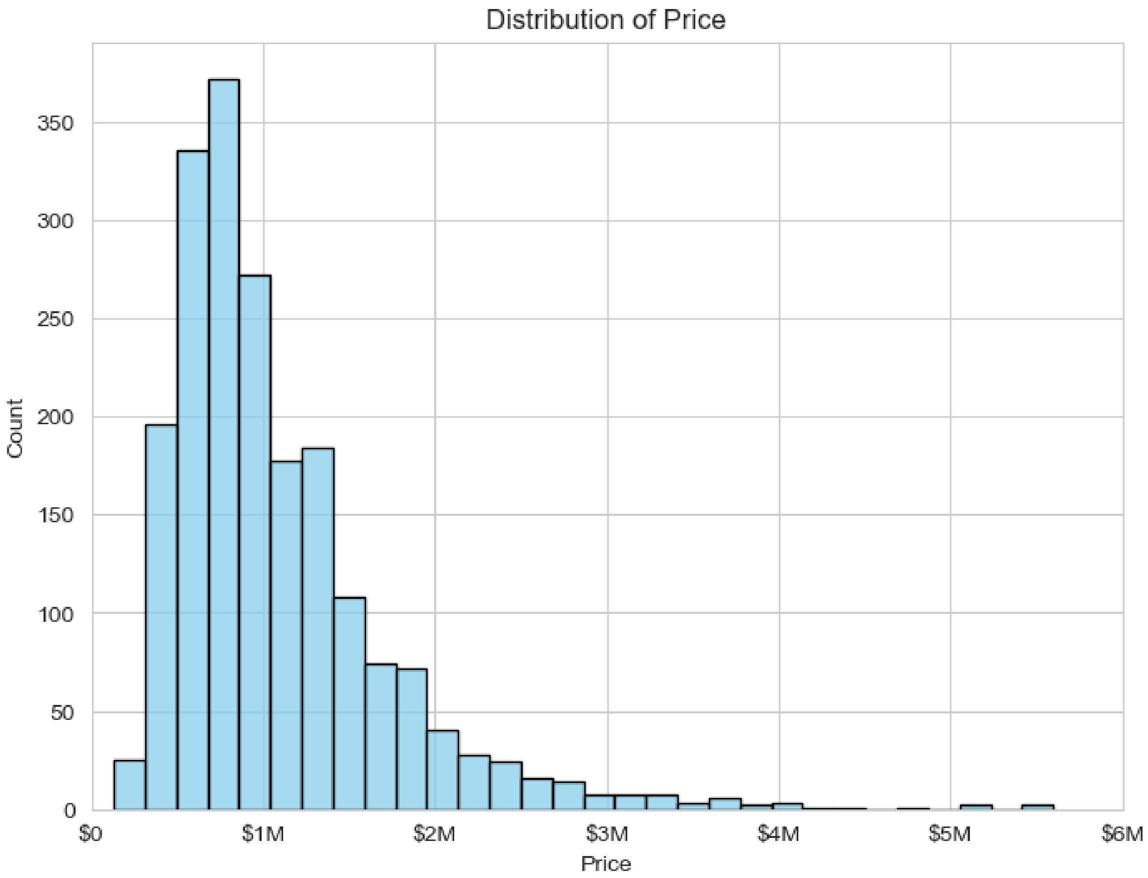
This count plot represents the distribution of car spots in properties, in descending order within the dataset. The majority of properties have 2 car spots: 973 counts, 49.2% proportion. In comparison, whilst 7 and 8 car spots in properties hold the same percentage at 0.1%, 7 is slightly worse with just a count of 1.



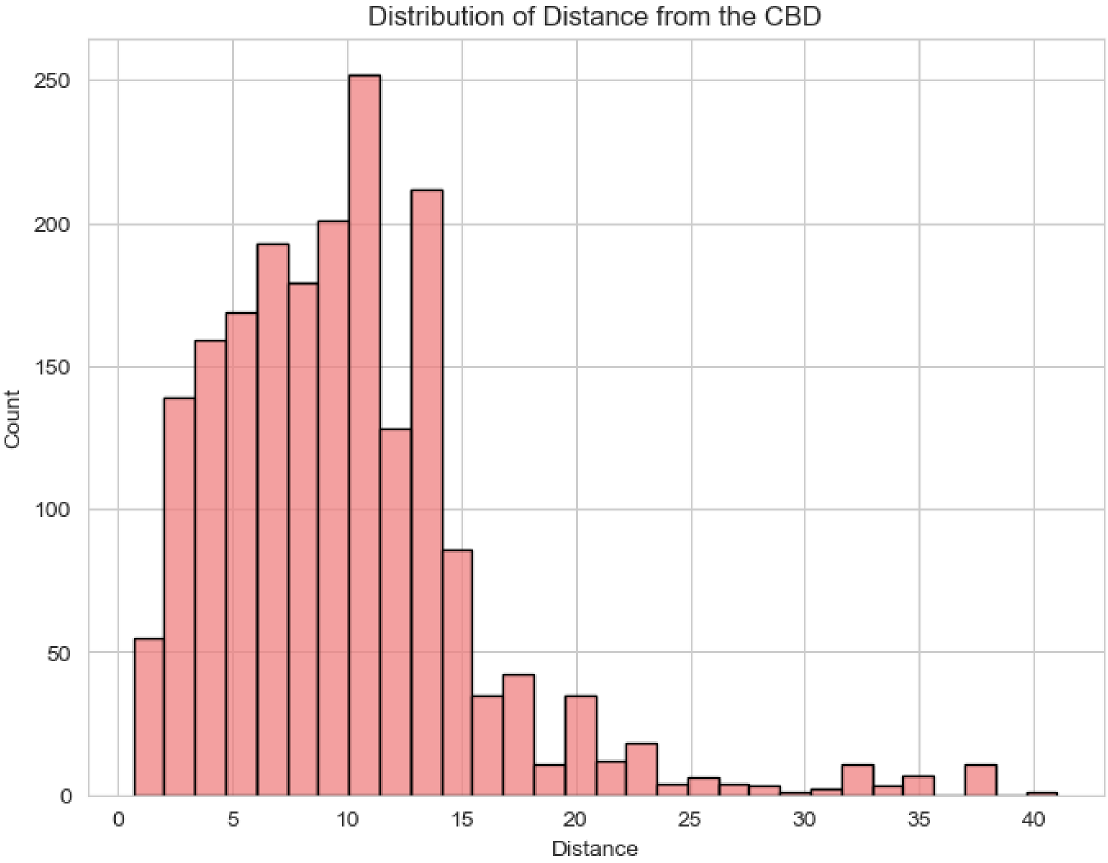
This count plot displays the council area distribution in descending order within the dataset. Moreland is by far the highest, having a count of 376 or proportionally 19% of the dataset. In contrast, Yarra Ranges is the lowest with a count of 1, representing a proportion of 0.1% within the dataset.



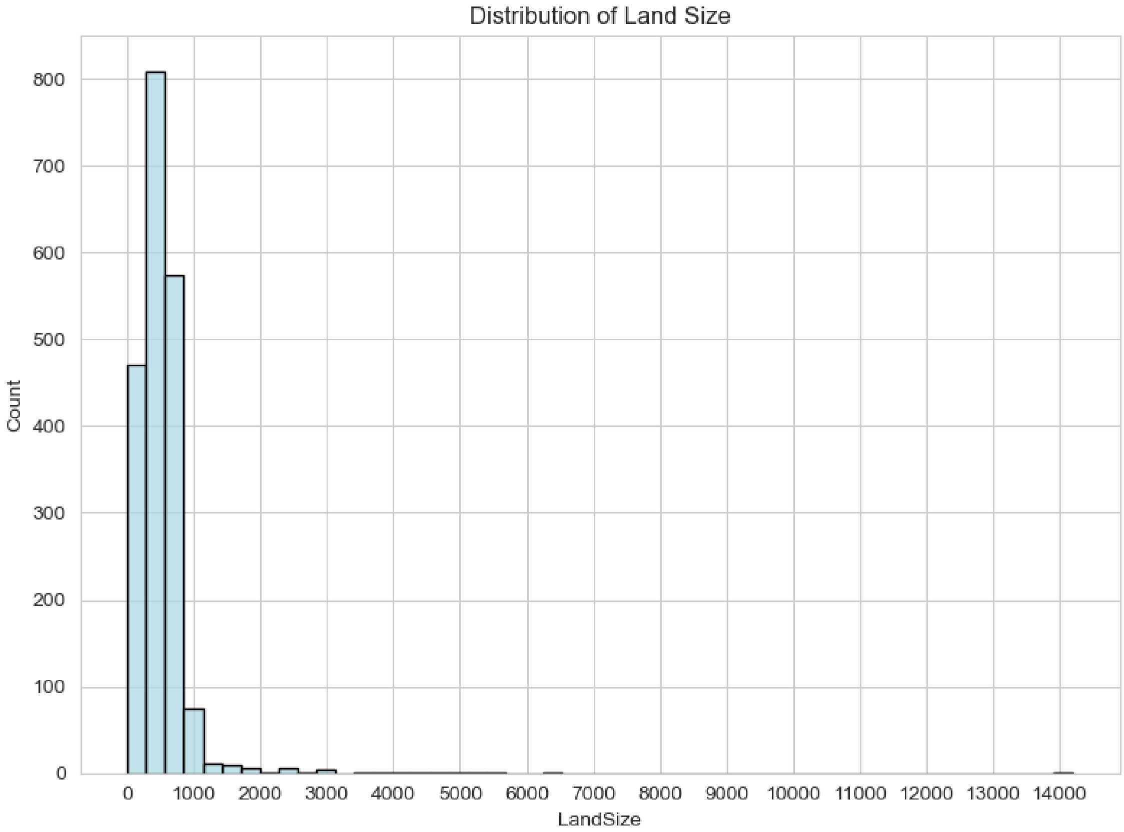
This count plot depicts the region distribution between properties within the dataset. The majority of properties fall under the Southern Metropolitan region: 681 counts, 34.4% of the data frame. The least amount of properties falls under the Western Victoria region: 1 count, 0.1% of the data frame.



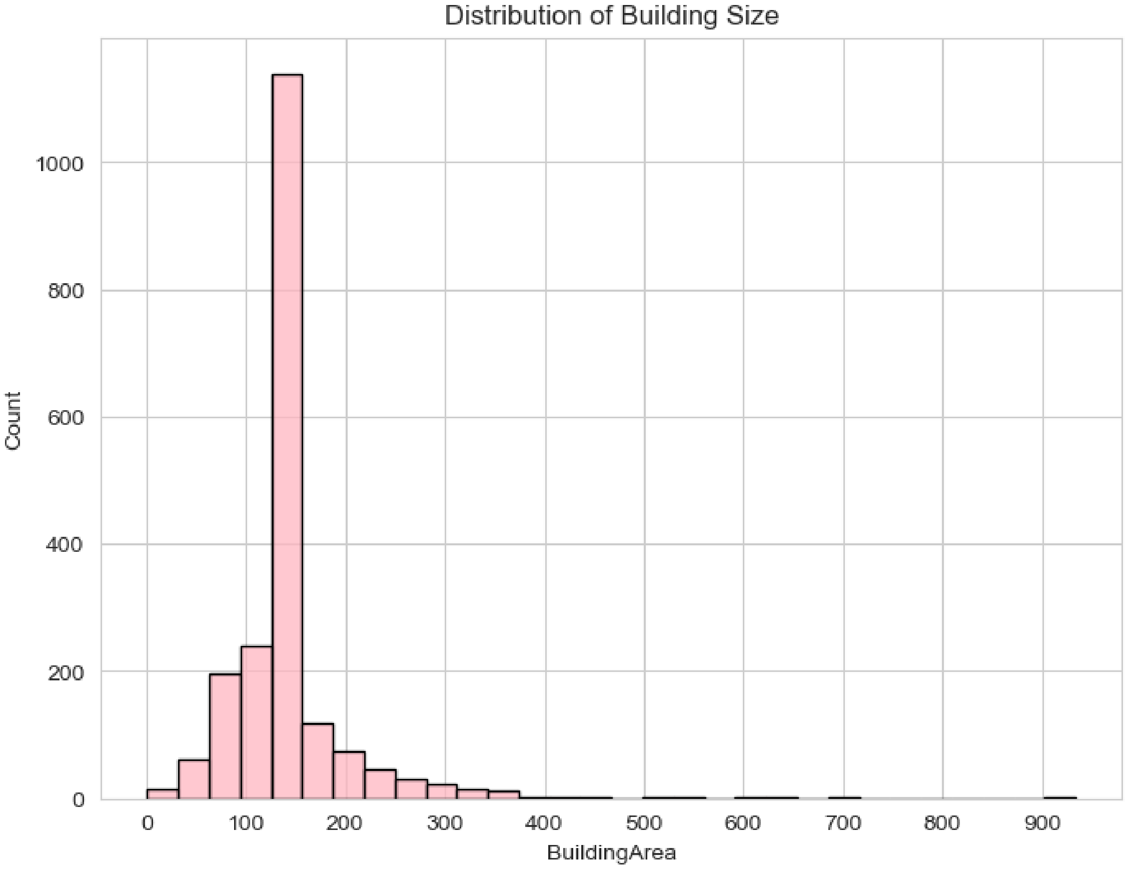
This histogram illustrates the distribution of the Price feature, which represents the prices of property within the dataset. The price range varies from 131,000 to $6,000,000. The most common price falls within the vicinity of $1,000,000, totaling approximately 375 counts. In contrast, the minority contingent is represented by valid outlier values, with values exceeding $4,000,000 and peaking at $5,600,000.



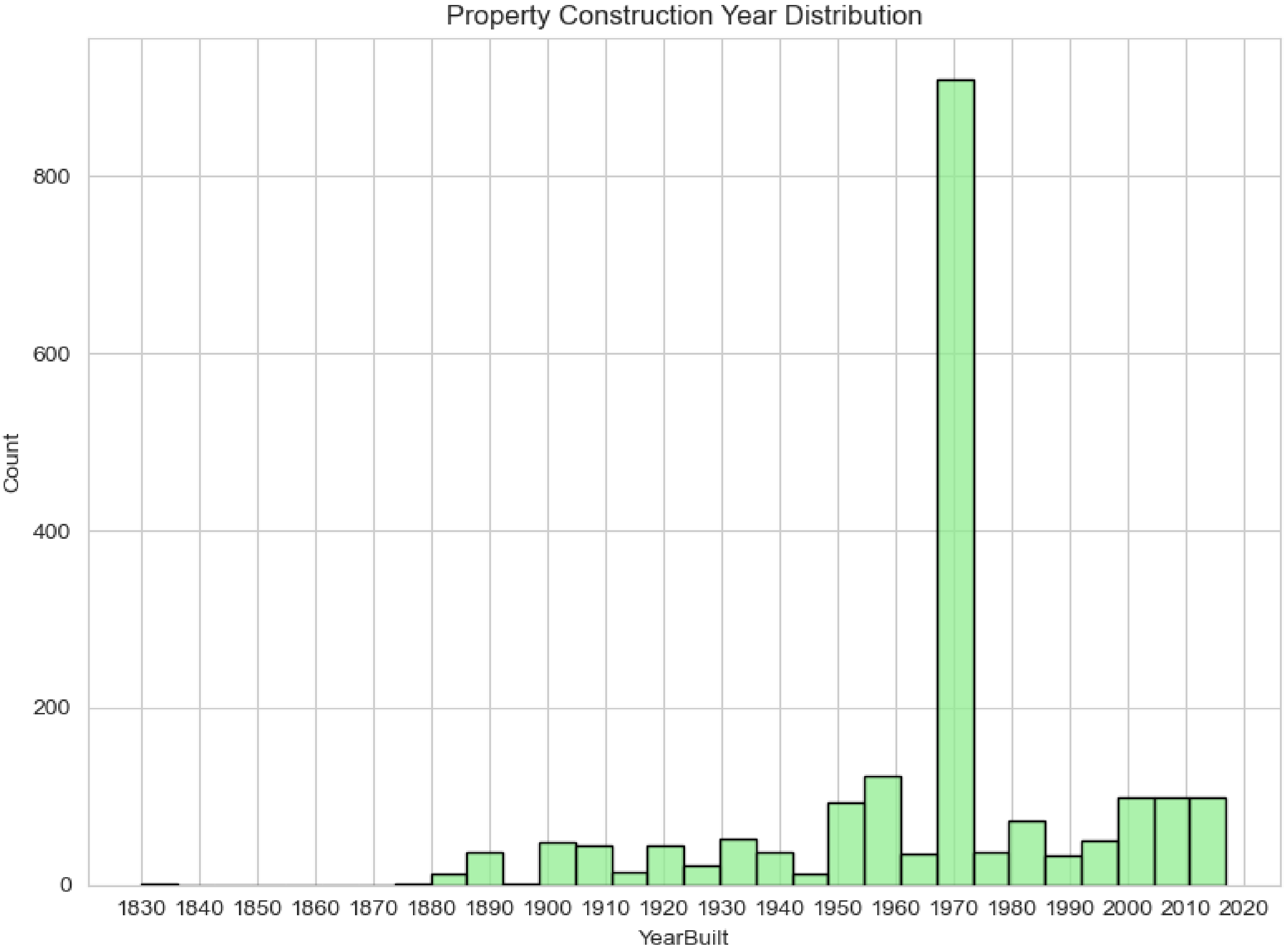
This histogram displays the distribution of distance (kilometers) from the CBD within the dataset. Distance from the CBD fluctuates between 0-41kms. The majority of properties are clustered around 12kms away from the CBD, accounting for roughly 250 observations. Conversely, the smallest subset of values are valid outliers, with values exceeding 30kms and peaking at 41kms from the CBD.



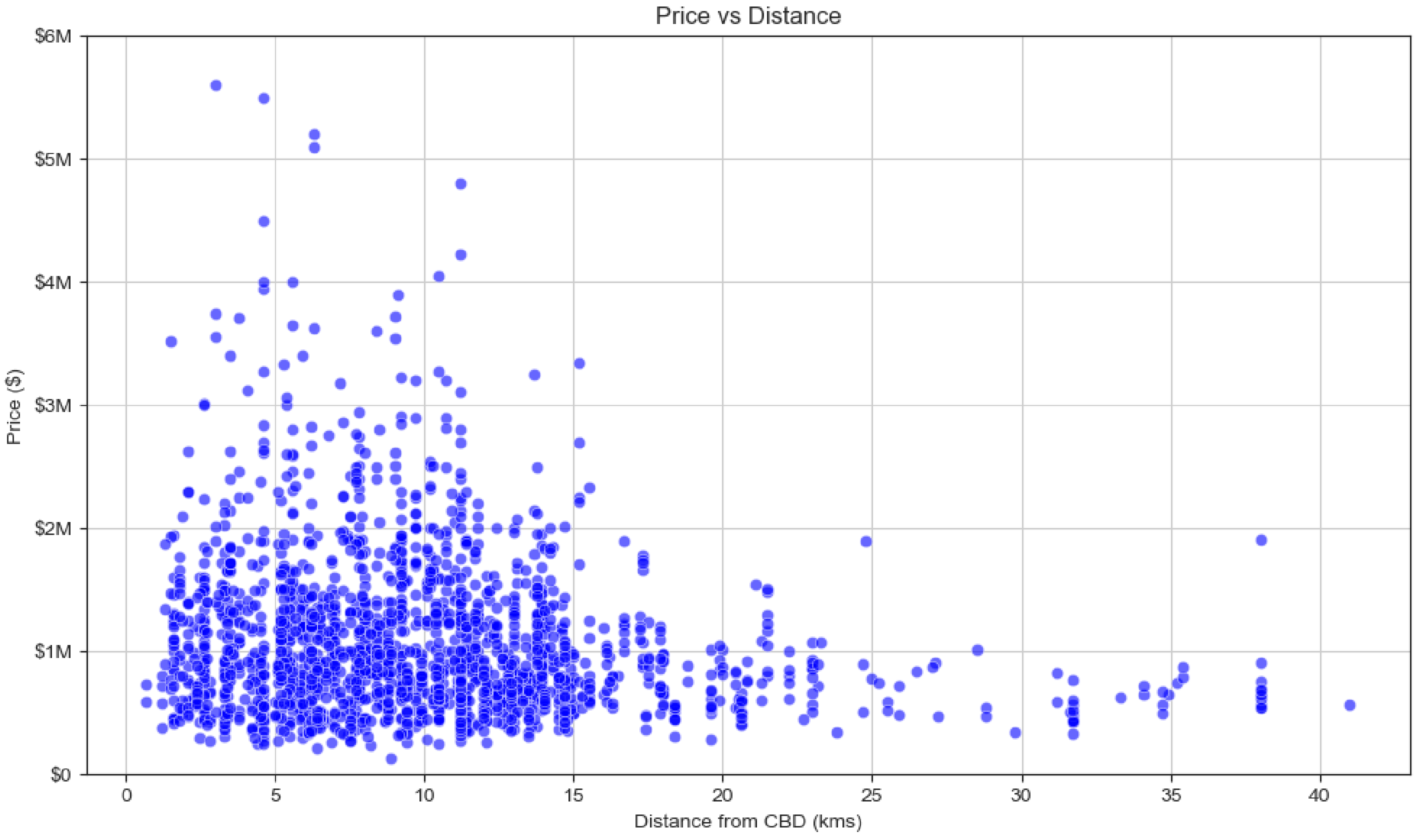
This histogram represents the distribution of the LandSize feature, which highlights the land size of all properties within the dataset, measured in square meters. Land size values range from 5-14196m^2. The most frequently occurring land size value is centered around 500m^2, accounting for approximately 810 observations. In contrast, the smallest subset of land size values are once again valid outliers: starting at 3500 and peaking at 14196m^2.



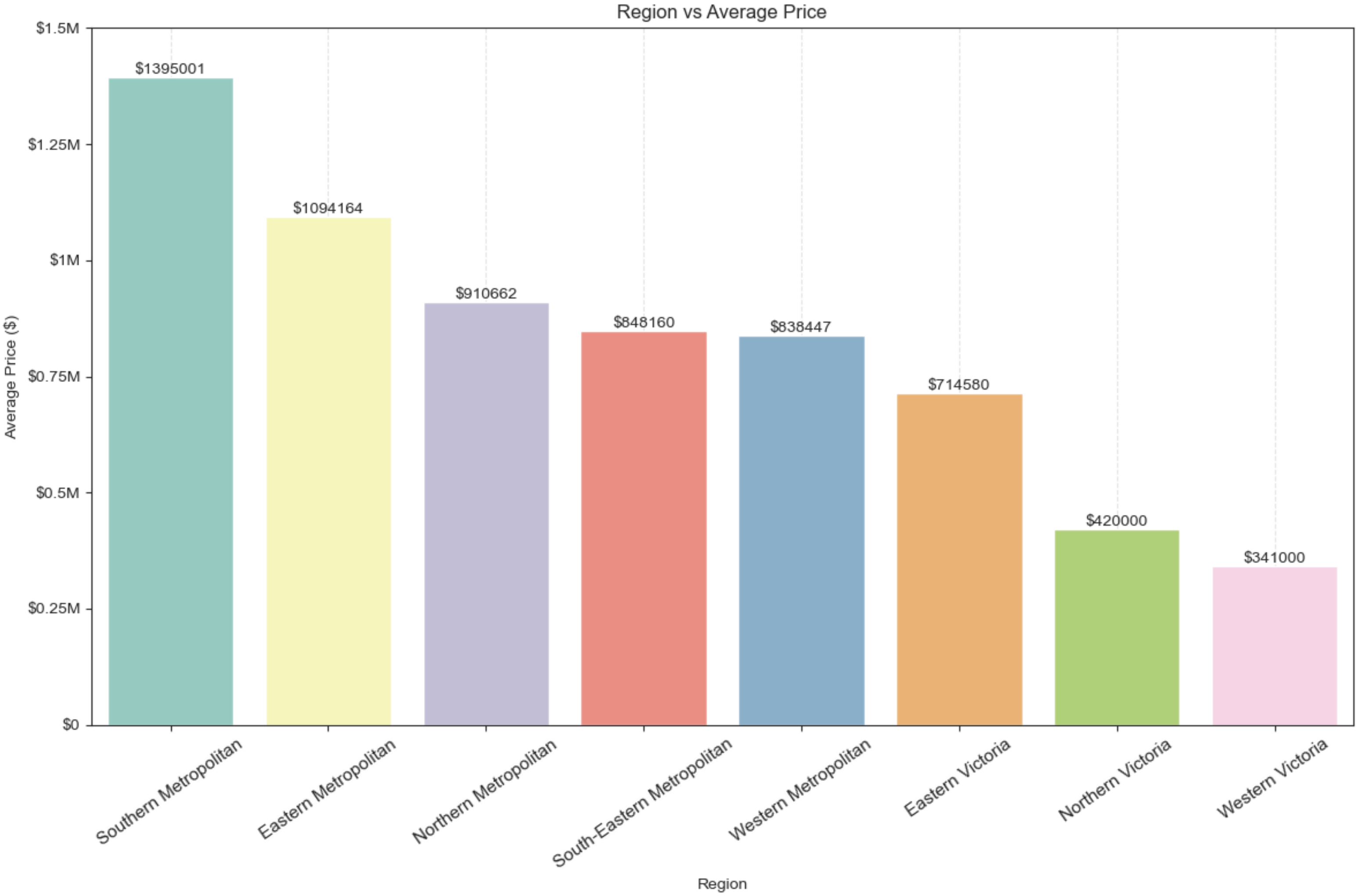
This histogram highlights the distribution of the BuildingArea feature, which is the measurement of the building size in every property within this dataset – measured in square meters. Building size values range from 1 to 934m^2.The most common building size is around 140m^2, containing approximately 1170 counts. In comparison, the valid outliers are the least frequently occurring values: 400-934m^2.



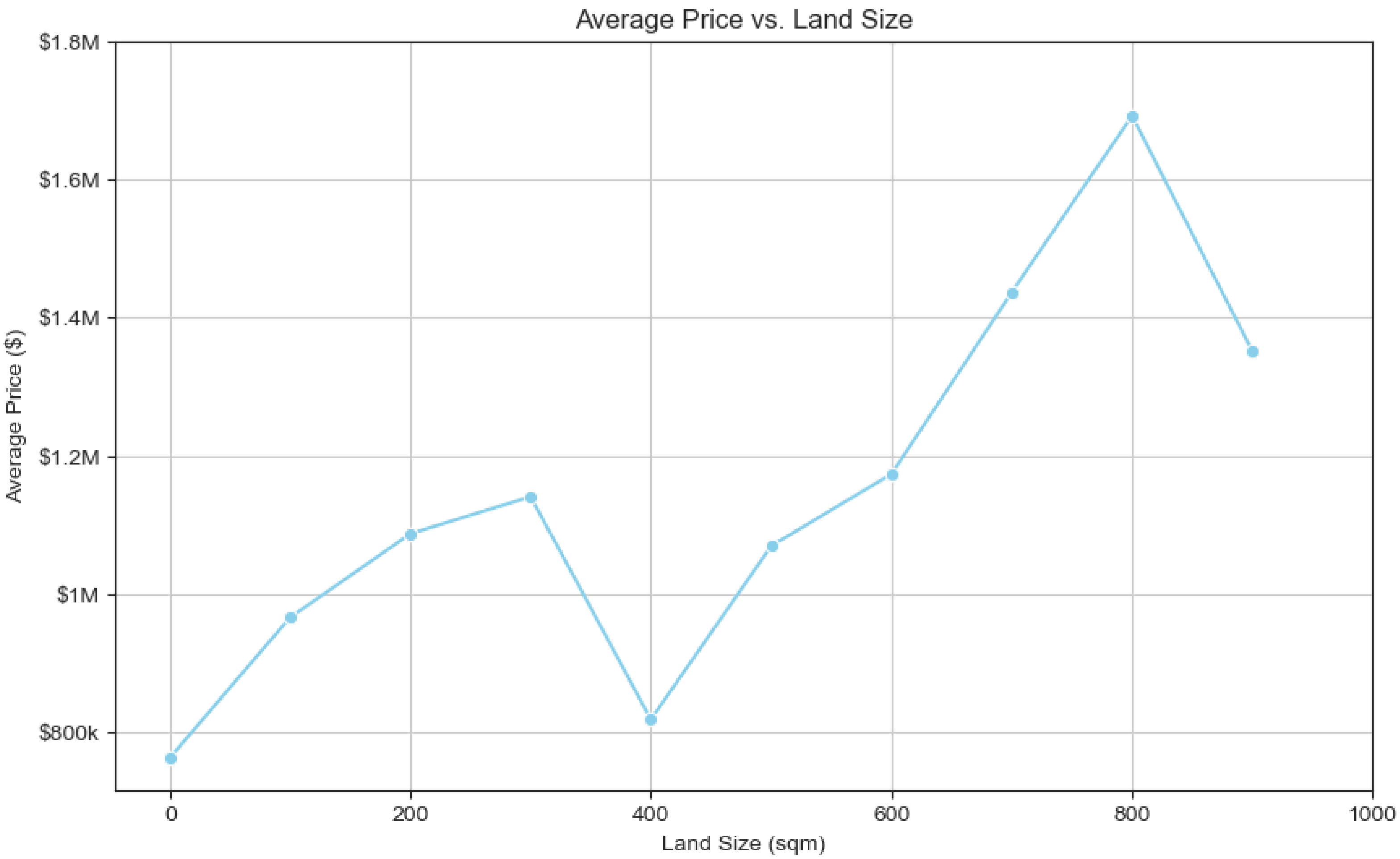
This histogram represents the distribution of the YearBuilt feature. Essentially, it highlights all the properties within the dataset and the corresponding years when they were built. The years range from 1830 to 2020. The most frequently occurring year is 1970, totaling a count of approximately 900. Conversely, the smallest frequency of values are the valid lower outliers (1830, 1877).



This scatterplot illustrates the relationship between property prices and distance from the CBD (measured in kilometers). It reveals a weak negative correlation: suggesting that as the distance from the CBD increases, there is a tendency for property prices to decrease. Conversely, properties located closer to the CBD tend to have higher prices.

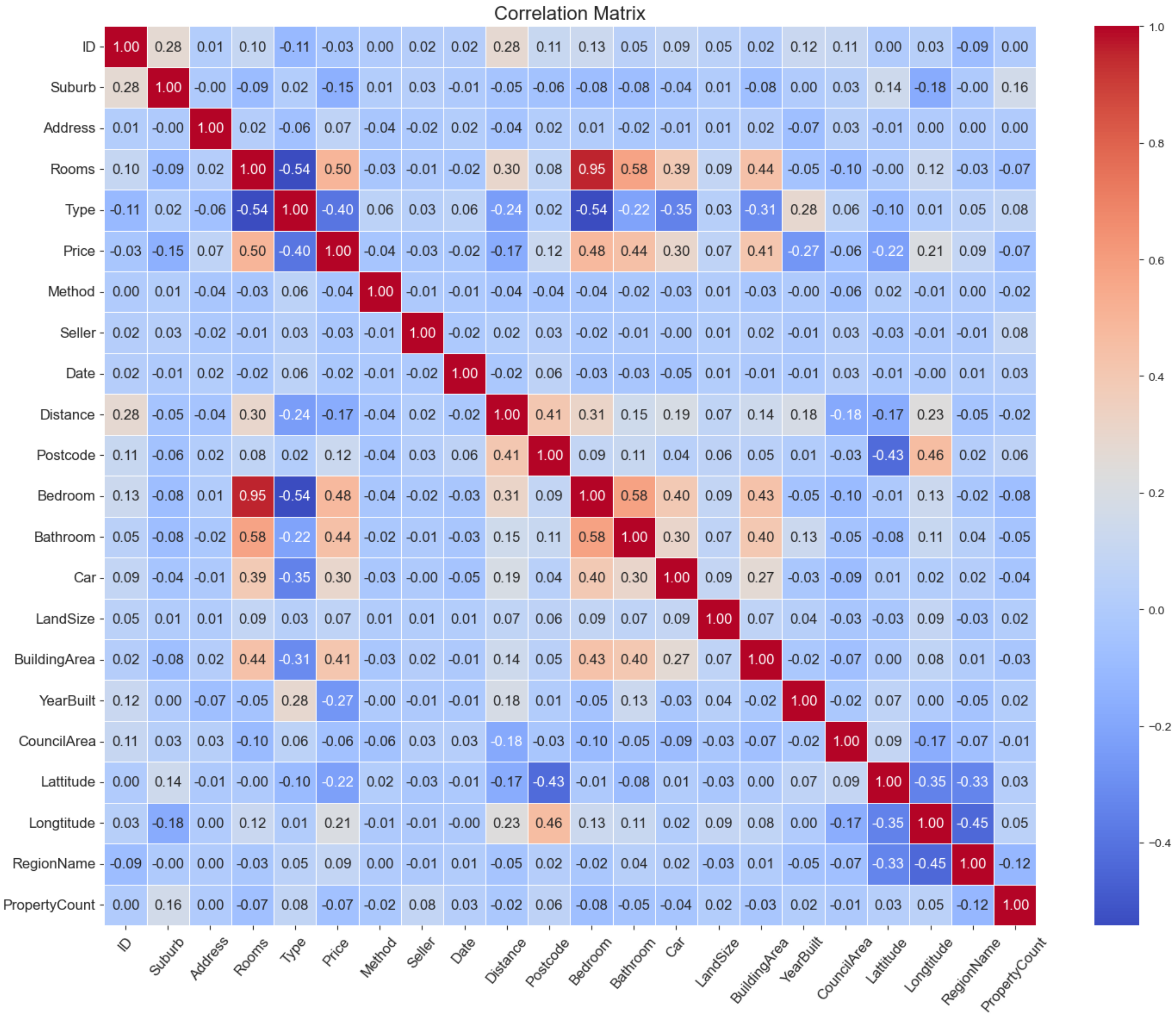


This bar plot illustrates the relationship between property prices and the different regions within the dataset. It highlights the average property prices across every single region. We can clearly observe that the Southern Metropolitan region has the highest average price of a property at $1,395,001. In sharp contrast, the Western Victoria region has the lowest average price of a property at $341,000.



This line chart depicts the relationship between Price and LandSize. Specifically, It shows the correlation between the average price vs land size (measured in square meters m^2) by displaying the relations between the average price of properties and their corresponding land size. By analyzing the trend of the line, we can observe the average price changing as the land size increases, and that there is a positive correlation between price and land size. Notably, at 400m^2 the line begins to sharply ascend and reaches its peak at 800m^2 - reaching approximately $1,700,000 at its peak. The price has more than doubled within this range (400-800m^2). This suggests that the average price of a property tends to significantly increase when land size increases. Overall, the observed trends indicate a positive correlation between the features Price and LandSize.

Task 2: Relationships Discovery Among Features



	ID	Suburb	Address	Rooms	Type	Price	Method	Seller	Date	Distance	Postcode	Bedroom	Bathroom	Car	LandSize	BuildingArea	YearBuilt	CouncilArea	Latitude	Longitude	RegionName	PropertyCount
ID	1.00	0.28	0.01	0.10	-0.11	-0.03	0.00	0.02	0.02	0.28	0.11	0.13	0.05	0.09	0.05	0.02	0.12	0.11	0.00	0.03	-0.09	0.00
Suburb	0.28	1.00	-0.00	-0.09	0.02	-0.15	0.01	0.03	-0.01	-0.05	-0.06	-0.08	-0.08	-0.04	0.01	-0.08	0.00	0.03	0.14	-0.18	-0.00	0.16
Address	0.01	-0.00	1.00	0.02	-0.06	0.07	-0.04	-0.02	0.02	-0.04	0.02	0.01	-0.02	-0.01	0.01	0.02	-0.07	0.03	-0.01	0.00	0.00	0.00
Rooms	0.10	-0.09	0.02	1.00	-0.54	0.50	-0.03	-0.01	-0.02	0.30	0.08	0.95	0.58	0.39	0.09	0.44	-0.05	-0.10	-0.00	0.12	-0.03	-0.07
Type	-0.11	0.02	-0.06	-0.54	1.00	-0.40	0.06	0.03	0.06	-0.24	0.02	-0.54	-0.22	-0.35	0.03	-0.31	0.28	0.06	-0.10	0.01	0.05	0.08
Price	-0.03	-0.15	0.07	0.50	-0.40	1.00	-0.04	-0.03	-0.02	-0.17	0.12	0.48	0.44	0.30	0.07	0.41	-0.27	-0.06	-0.22	0.21	0.09	-0.07
Method	0.00	0.01	-0.04	-0.03	0.06	-0.04	1.00	-0.01	-0.01	-0.04	-0.04	-0.04	-0.02	-0.03	0.01	-0.03	-0.00	-0.06	0.02	-0.01	0.00	-0.02
Seller	0.02	0.03	-0.02	-0.01	0.03	-0.03	-0.01	1.00	-0.02	0.02	0.03	-0.02	-0.01	-0.00	0.01	0.02	-0.01	0.03	-0.03	-0.01	-0.01	0.08
Date	0.02	-0.01	0.02	-0.02	0.06	-0.02	-0.01	-0.02	1.00	-0.02	0.06	-0.03	-0.03	-0.05	0.01	-0.01	-0.01	0.03	-0.01	-0.00	0.01	0.03
Distance	0.28	-0.05	-0.04	0.30	-0.24	-0.17	-0.04	0.02	-0.02	1.00	0.41	0.31	0.15	0.19	0.07	0.14	0.18	-0.18	-0.17	0.23	-0.05	-0.02
Postcode	0.11	-0.06	0.02	0.08	0.02	0.12	-0.04	0.03	0.06	0.41	1.00	0.09	0.11	0.04	0.06	0.05	0.01	-0.03	-0.43	0.46	0.02	0.06
Bedroom	0.13	-0.08	0.01	0.95	-0.54	0.48	-0.04	-0.02	-0.03	0.31	0.09	1.00	0.58	0.40	0.09	0.43	-0.05	-0.10	-0.01	0.13	-0.02	-0.08
Bathroom	0.05	-0.08	-0.02	0.58	-0.22	0.44	-0.02	-0.01	-0.03	0.15	0.11	0.58	1.00	0.30	0.07	0.40	0.13	-0.05	-0.08	0.11	0.04	-0.05
Car	0.09	-0.04	-0.01	0.39	-0.35	0.30	-0.03	-0.00	-0.05	0.19	0.04	0.40	0.30	1.00	0.09	0.27	-0.03	-0.09	0.01	0.02	0.02	-0.04
LandSize	0.05	0.01	0.01	0.09	0.03	0.07	0.01	0.01	0.01	0.07	0.06	0.09	0.07	0.09	1.00	0.07	0.04	-0.03	-0.03	0.09	-0.03	0.02
BuildingArea	0.02	-0.08	0.02	0.44	-0.31	0.41	-0.03	0.02	-0.01	0.14	0.05	0.43	0.40	0.27	0.07	1.00	-0.02	-0.07	0.00	0.08	0.01	-0.03
YearBuilt	0.12	0.00	-0.07	-0.05	0.28	-0.27	-0.00	-0.01	-0.01	0.18	0.01	-0.05	0.13	-0.03	0.04	-0.02	1.00	-0.02	0.07	0.00	-0.05	0.02
CouncilArea	0.11	0.03	0.03	-0.10	0.06	-0.06	-0.06	0.03	0.03	-0.18	-0.03	-0.10	-0.05	-0.09	-0.03	-0.07	-0.02	1.00	0.09	-0.17	-0.07	-0.01
Lattitude	0.00	0.14	-0.01	-0.00	-0.10	-0.22	0.02	-0.03	-0.01	-0.17	-0.43	-0.01	-0.08	0.01	-0.03	0.00	0.07	0.09	1.00	-0.35	-0.33	0.03
Longtitude	0.03	-0.18	0.00	0.12	0.01	0.21	-0.01	-0.01	-0.00	0.23	0.46	0.13	0.11	0.02	0.09	0.08	0.00	-0.17	-0.35	1.00	-0.45	0.05
RegionName	-0.09	-0.00	0.00	-0.03	0.05	0.09	0.00	-0.01	0.01	-0.05	0.02	-0.02	0.04	0.02	-0.03	0.01	-0.05	-0.07	-0.33	-0.45	1.00	-0.12
PropertyCount	0.00	0.16	0.00	-0.07	0.08	-0.07	-0.02	0.08	0.03	-0.02	0.06	-0.08	-0.05	-0.04	0.02	-0.03	0.02	-0.01	0.03	0.05	-0.12	1.00

This is a correlation matrix between all features within the dataset. In every row you can observe said features correlation against every other feature. The closer the number is to 1 (or -1), the stronger the positive or negative correlation is. We wont go through every feature here as there is a lot, however we will target one of the most important features in this dataset - Price.

- Rooms, Bedroom, Bathroom, BuildingArea, Car, Longitude; have a positive correlation with Price – with rooms having the highest rate of positive correlation at 0.50.
- Address, Postcode, LandSize, RegionName; have a weak/low positive correlation with Price.

- ID, Method, Seller, and Date have no correlation with Price.
- CouncilArea, PropertyCount; have a weak/low negative correlation with Price.
- Type, YearBuilt, Latitude, Distance, Suburb; have a negative correlation with Price – with Type having the highest rate of negative correlation at -0.40.

Task 3: List Any Potential Business Analysis Tasks With Your Justification

Business Analysis Tasks:

1. **Association Rule Mining:** Association Rule Mining excels at identifying hidden patterns and relationships between features in a dataset. Particularly, in finding which features are together when a customer purchases a product or service and what their corresponding relationship is. This method can be utilized at our real estate consulting firm with the provided Melbourne housing dataset in many ways, for example to discover key insights into customer preferences and behavior. E.g., it could unveil that properties that have a higher number of bedrooms are often associated with larger land size, bathrooms, and higher prices: with these insights in mind, our real estate consulting firm can tailor its marketing strategies to target specific customers more successfully. By highlighting properties with desirable features, such as multiple bedrooms and car spots, big land sizes, or a higher number of bathrooms; the firm can target and attract multiple potential buyers or renters who are more likely to be captivated by such properties. This personalized approach can lead to higher customer satisfaction, increased engagement, and ultimately, more successful transactions.

   - **Associate Rule Mining Business Benefits:**
     - Understanding the correlations between features facilitates the targeting of marketing campaigns towards specific individuals and the customization of property recommendations to suit individual preferences.
     - Guides pricing strategies and identifies opportunities for promoting additional services or related products based on customer behavior/ preferences.
     - By uncovering hidden patterns, association analysis assists in optimizing property listings and improving overall client satisfaction.

2. **Classification:** Classification is mainly used to build predictive models to classify a certain label (target feature) and evaluate itself on how accurate its predictions are. In the context of our real estate consulting firm and the provided Melbourne housing dataset, we could predict/classify a lot of labels. For example, we can build predictive models that classify properties on whether or not it will be sold swiftly, based on the sales method, seller, rooms, postcode, and property type. The insights gathered from the predictive models could be utilized to target specific demographics more effectively. The main benefit with classification is its predictive model capabilities and how good the model we build actually is, as it will be tested on unseen data,

simulating real world scenarios were it predicts whether a property will be sold swiftly or not based on its features.

- **Classification Business Benefits:**
  - Predictive modelling aids in identifying properties with unique features that appeal to specific buyers or renters.
  - Classifying/predicting properties allows our real estate consulting firm to tailor marketing strategies, advertisements, and client interactions to target the right audience.
  - These predictive models can also assist in optimizing inventory management and resource allocation by classifying demand for different property types and price ranges accurately.

3. **Regression:** Regression would mainly be used here to predict property prices based on multiple features within the provided Melbourne housing dataset. For example, regression analysis would be used to predict the price of a property based on its features such as the number of rooms, distance from CBD, land size, building area, etc. By feeding regression all of the associated features of properties within the housing dataset, it would identify patterns in property pricing over time and predict future price movements; the main benefit with this is the valuable key insights you could extract and in turn make informed decisions about property investments. Additionally, you could utilize these key insights to identify what exact features influence property prices the most: this information is also beneficial to the cause of property valuation and pricing strategies. Ultimately, regression analysis allows our real estate consulting firm to maximize returns.

- **Regression Business Benefits:**
  - Regression property price prediction models provide valuable insights for buyers and sellers, facilitating informed investment decisions with accurate estimations of property values.
  - By analyzing our data, regression models reveal past price patterns and forecast future market shifts, empowering our real estate consulting firm to foresee changes and adjust strategies accordingly.
  - Regression analysis identifies the most influential features impacting property prices, allowing our real estate consulting firm to focus on optimizing these areas to boost returns and profitability.

In summary, leveraging data mining methods like association, classification, and regression on our Melbourne housing dataset will provide numerous valuable insights into customer behaviour, market trends, and property valuation. By harnessing these insights, our real estate consulting firm can substantially enhance their strategic planning, support decision-making processes, and ultimately drive business growth and profitability.

Conclusion

Business Analytics serves as an extremely powerful tool for real estate consulting firms, equipping them with an extensive array of advanced techniques and tools, including data analysis and data mining methods. Through these sophisticated approaches, the firm can uncover key patterns, trends, and relationships in the data, which would otherwise remain hidden. These key insights garnered from business analytics have the potential to exponentially enhance the firm's operations - driving a plethora of significant business benefits. By harnessing the power of business analytics, the real estate consulting firm gains a competitive edge in the market. With access to actionable insights derived from comprehensive data analysis, the firm can make informed decisions, optimize strategies, and capitalize on emerging opportunities. This distinct advantage is particularly crucial in a fiercely competitive market landscape, where staying ahead of the curve is paramount. In conclusion, leveraging business analytics on the housing dataset provided empowers the real estate consulting firm to unlock invaluable insights, steering substantial improvements in overall efficiency. By embracing data-driven decision-making, the firm can navigate complexities with confidence and chart a path towards sustained success in the real estate industry.