Property Data Analysis

Toby Ryu

Summer 2025

About the data

The data has four columns:

- PropertyID
  - ID number for the property ranging from 2001-2100
- Location
  - Either downtown, suburb, or countryside
- Size_sqft
  - The property size in square feet
- SalePrice
  - The price of the property

```
[11]: properties_df
```

[11]:

	PropertyID	Location	Size_sqft	SalePrice
0	2001	Suburb	2996	355372
1	2002	Downtown	2887	505830
2	2003	Suburb	957	338789
3	2004	Countryside	1814	419335
4	2005	Suburb	850	357301
...	...	...	...	...
95	2096	Countryside	2005	271942
96	2097	Suburb	2686	551317
97	2098	Countryside	1889	530264
98	2099	Downtown	1453	502089
99	2100	Countryside	2050	562155

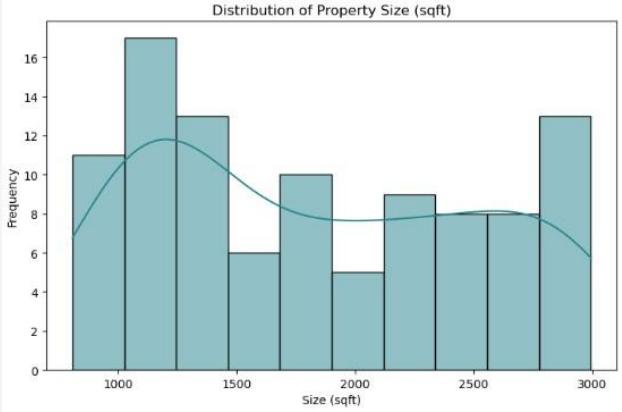
100 rows × 4 columns

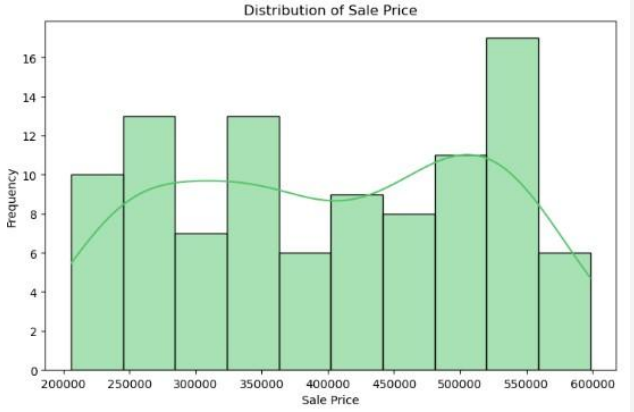
Let's get started. Our goal is to look for patterns in the data that we can investigate. We will analyze visualizations made by Matplotlib and Seaborn.

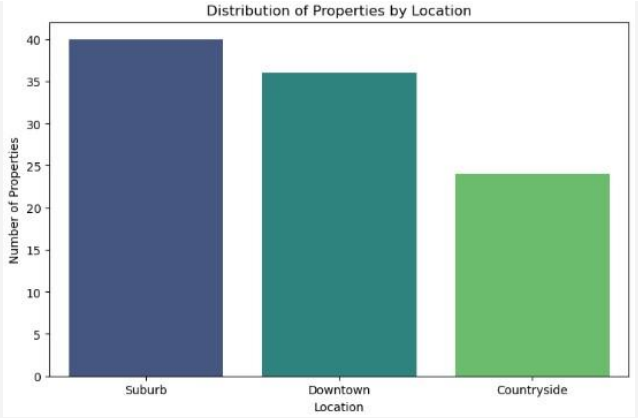
Distributions

Property size and sale price often have right-skewed distributions. This is due to majority of the population living in affordable housing in contrast with extremely wealthy property owners. Our distributions, however, are more uniform than anything and don't seem to have any outliers. We will hold off on any assumptions until viewing their linear relationship.

Our location distribution reveals unequal group sizes but not alarmingly. We can use this to our advantage to compare price and size by location.







Relationship Analysis

The scatter plot reveals no linear relationship between size and price as well as no clear clusters. This means neither size nor location are able to predict property size.

The box plots reveal how similar sale prices are between locations. In addition to having similar ranges, each location also has similar interquartile ranges.

Correlation Coefficient

As expected, our correlation heatmap reveals a very low correlation coefficient between property size and sale price. -0.031 leads us to conclude that a linear relationship between the two variables is nonexistent, eliminating any possibility for a predictive model.

Analysis

Initially, we hoped to see if either property size or location could explain property sale price.

The scatter plot revealed no patterns, the correlation coefficient revealed very low correlation, and the box plots revealed evenly distributed data.

We can conclude that our variables have no predictive power for sale price.

Our findings tell us that unobserved factors like property condition, property age, proximity to schools, or amenities could have more predictive power.

Our next steps should be to collect these additional variables.

We can also narrow our focus to a single location, such as downtown or the suburbs, to better capture underlying patterns.

Next Steps

Thank you

Questions? Contact
tobyhryu@gmail.com