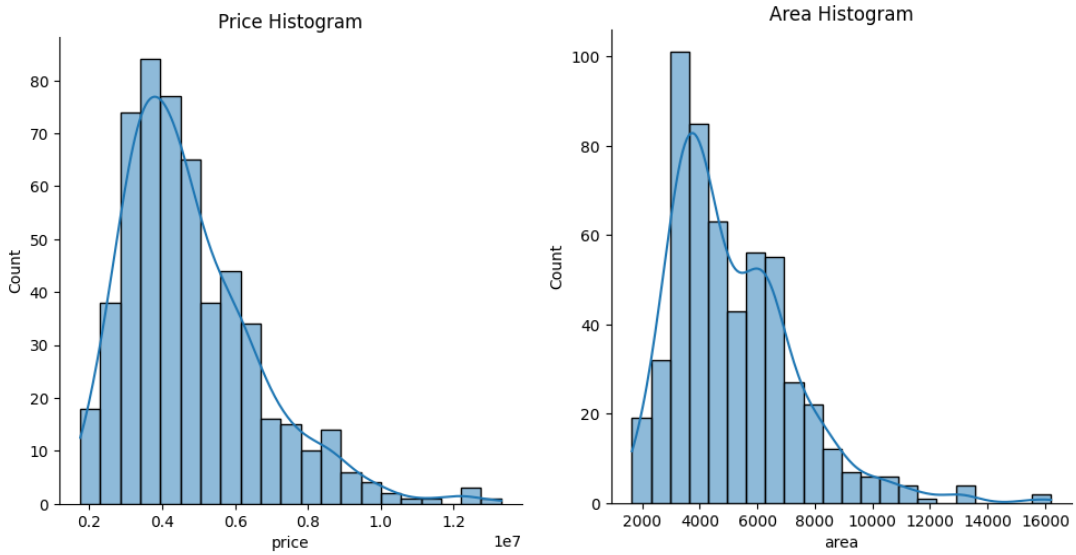
DATA1030 Mid Term Report

1. Introduction:

   In this project, we will be exploring the housing prices dataset sourced from Kaggle. We are given 12 features, which are the following: area, number of bedrooms, number of bathrooms, stories, main road, guest room availability, basement, hot water heating, air conditioning, parking, furnishing status and pref. area. We have 545 available data points.

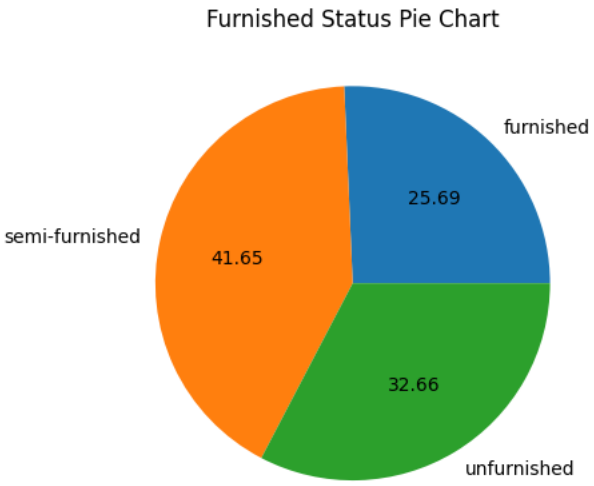
   The target variable in this case is the price of the house. As we are attempting to predict the price of the house, this is a regression task. I believe that this is an interesting dataset to look at for various reasons. Firstly, housing price is the topic of many contentious reasons. For some, it is an investment that they need to get right as it is a pricy asset, while for those looking to purchase, they do not want their prices to decrease over the years of their residence.

2. Exploratory Data Analysis



   As we look at the distribution of the house, we can see that there is a slight skew, but this is due to the few extreme outliers that occur as a result of the luxury houses present in the dataset (houses with 6+ bedrooms, etc). If we exclude those data points, the distribution of the prices would be roughly normally distributed. It is interesting to see from the histogram of the area that the area is also distributed in a similar fashion. This makes sense as the price and the area should roughly be linearly correlated in my view.

3. In the data pre-processing pipeline, I split the data into 80% training and 20% testing and it is also shuffled. The dataset is IID.



Another interesting graphic to look at is the furnished status of the houses. About a quarter of the houses come furnished, third unfurnished and the rest semi furnished.