Main Project Stage 1

Data Source

King's County Housing Data

Sourced from Kaggle: [https://www.kaggle.com/datasets/harlfoxem/housesalesprediction]

It is under a CC0: Public Domain license.

The dataset is for housing in King's Country sold between May 2014 and May 2015, with metrics for each individual house sale.

Cleaning

Not too much cleaning had to be done. I removed data which I felt was unusable (the specific ID of the sale and the zip code). I moved the price of the house to the first column, and I also cleaned up the date of sale so that it was numeric.

Visualization 1: Histograms of Some Housing Attributes

I showed the bedroom count, bathroom count, and floor count in a histogram. You can see the general distributions of how many bedrooms/bathrooms/floors each house has. Note that the bedroom count histogram appears to go up to 20+: this is because there's a single house at element 15870 with 33 bedrooms. I chose not to clean this because I do not know the true bedroom count, but I am noting that this is likely an error in the data.

Visualization 2: House Sale Locations

These are the locations of house sales in the dataset. The reason I did not clean out the location data is because, as seen in the visualization, it forms coherent shapes that may help the neural network.

Visualization 3: Histogram of House Price (in millions)

Thsi is the distribution of house prices, with the x-axis labeled in millions of dollars. The neural network will predict housing prices, so this is the distribution of the ouput parameter.