Problem Statement Worksheet (Hypothesis Formation)

How can I create a model that is able to predict New York property sales prices between 2022-2023 to within 20% margin of error based on intrinsic features of the property such neighbourhood, building type, square footage?

H

1 Context

The Department of Finance's Rolling Sales files lists properties that sold in during 2022- 2023 in New York City for residential units, co-operative and condominiums, offices and warehouses. These files include features such as the neighborhood, building type, square footage, sales data and other categorical and numerical features.

2 Criteria for success

Creating a prediction model should strive to produce a sales price within 20% margin of error.

3 Scope of solution space

The model will only be limited to intrinsic features of the property and not include macro economic factors. The data set is limited to the 5 boroughs of New York City and only trained on sales data from 2022-2023.

4 Constraints within solution space

The data set has many null and empty values that requires a complete understanding of the data set before cleaning can be completed.

5 Stakeholders to provide key insight

Springboard Mentor – Branko Kovac

6 Key data sources

https://www.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

The data webpage has 5 MS EXCEL worksheets for 5 Boroughs that needs to be combined to tackle this problem. The combined data will needed to be split into 70% training and 30% testing

H D E I P