Idea 2: NYC Property Sales Data. Investing in real estate is very similar to investing in many other financial instruments. Breaking down the price data based on attributes is akin to factor analysis in financial modeling. Although not financial company data, this is a good data set with translatable skills for my field of study.

Data Features:

5 XLSX tables (across 5 NYC boroughs) which need to be combined to appended together for the whole project.

Each excel dataset has 20 columns describing the sale observation, with "SALES PRICE" being the target column. Each excel sheet has about 20,000 rows of observations. Combining all 5 will give roughly 100,000 sales observations.

The columns / features are described here by: <https://www.nyc.gov/site/finance/taxes/glossary-property-sales.page>

Out of the columns, there are 7 numerical features and 13 categorical features.

Complications:

Data has many incomplete entries and 0 observations that need to be cleaned before any sort of classification can be performed.

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

Github: https://github.com/jcheng93/Capstone-NYC-Property-Sales