1 Part 1: Contextualizing the Data

Let's try to understand the background of our dataset before diving into a full-scale analysis.

1.1 Question 1a

Based on the columns present in this data set and the values that they take, what do you think each row represents? That is, what is the granularity of this data set?

One property in Cook County IL

1.2 Question 1b

Why do you think this data was collected? For what purposes? By whom?

This question calls for your speculation and is looking for thoughtfulness, not correctness.

Collected by Cook County Department of Housing or Urban Development to follow trends in property development and how different changes can affect, whether positively or negatively, the overall value of the properties. With this information, users can find average values for different townships, neighborhoods, property types, etc. and can look for what changes have the best net effect on the property's value. Development companies and their data scientists could use this information to determine whether or not their development projects are financially worthwhile ventures.

1.3 Question 1c

Craft at least two questions about housing in Cook County that can be answered with this data set and provide the type of analytical tool you would use to answer it (e.g. “I would create a ____ plot of ____ and ***” or ”I would calculate the*** [summary statistic] for ____ and ______”). Be sure to reference the columns that you would use and any additional data sets you would need to answer that question.

1) How does proximity to O’Hare Airport impact property values in Cook County?

I would perform regression analysis with Sale Price as the dependent variable and include O’Hare Noise as the independent variable to evaluate its impact. I would need to control for things like Lot Size, Age, Central Air, Central Heating, Garage, Building Square Feet, Neighborhood Code, and Property Class. I would need to use the Latitude and Longitude to account for spatial effects.

2) How do characteristics of the property affect the sale price in various neighborhoods in Cook County?

I would perform a Multivariate Analysis of Variance with Sale Price as the dependent variable again. I would include Neighborhood Code to analyze differences across neighborhoods. I would consider Land Square feet, Building square feet, Age, Central Air, Central Heating, Garage, Roof Material, Basement Finish, Fireplaces, Attic Type, Construction Quality, and Site Desirability as independent variables to assess how these characteristics impact Sale Price for different Neighborhood Code categories.

1.4 Question 1d

Suppose now, in addition to the information already contained in the dataset, you also have access to several new columns containing demographic data about the owner, including race/ethnicity, gender, age, annual income, and occupation. Provide one new question about housing in Cook County that can be answered using at least one column of demographic data and at least one column of existing data and provide the type of analytical tool you would use to answer it.

How does demographic profile of property owners (race/ethnicity, gender, age, income) influence Sale Prices in Cook County?

I would use Multilevel Regression analysis with Sale Price as the dependent variable and Owner Race/Ethnicity, Owner Gender, Owner Age, Owner Annual Income, and Owner Occupation as the independent variables while controlling for characteristics like Building Square Feet, Land Square Feet, Age, Garage, Central Air/Heating, Estimate (Land), and Estimate (Building). I would again need to consider Neighborhood Code to account for variations between different neighborhoods.

1.5 Question 2a

Using the plots above and the descriptive statistics from `training_data['Sale Price'].describe()` in the cells above, identify one issue with the visualization above and briefly describe one way to overcome it.

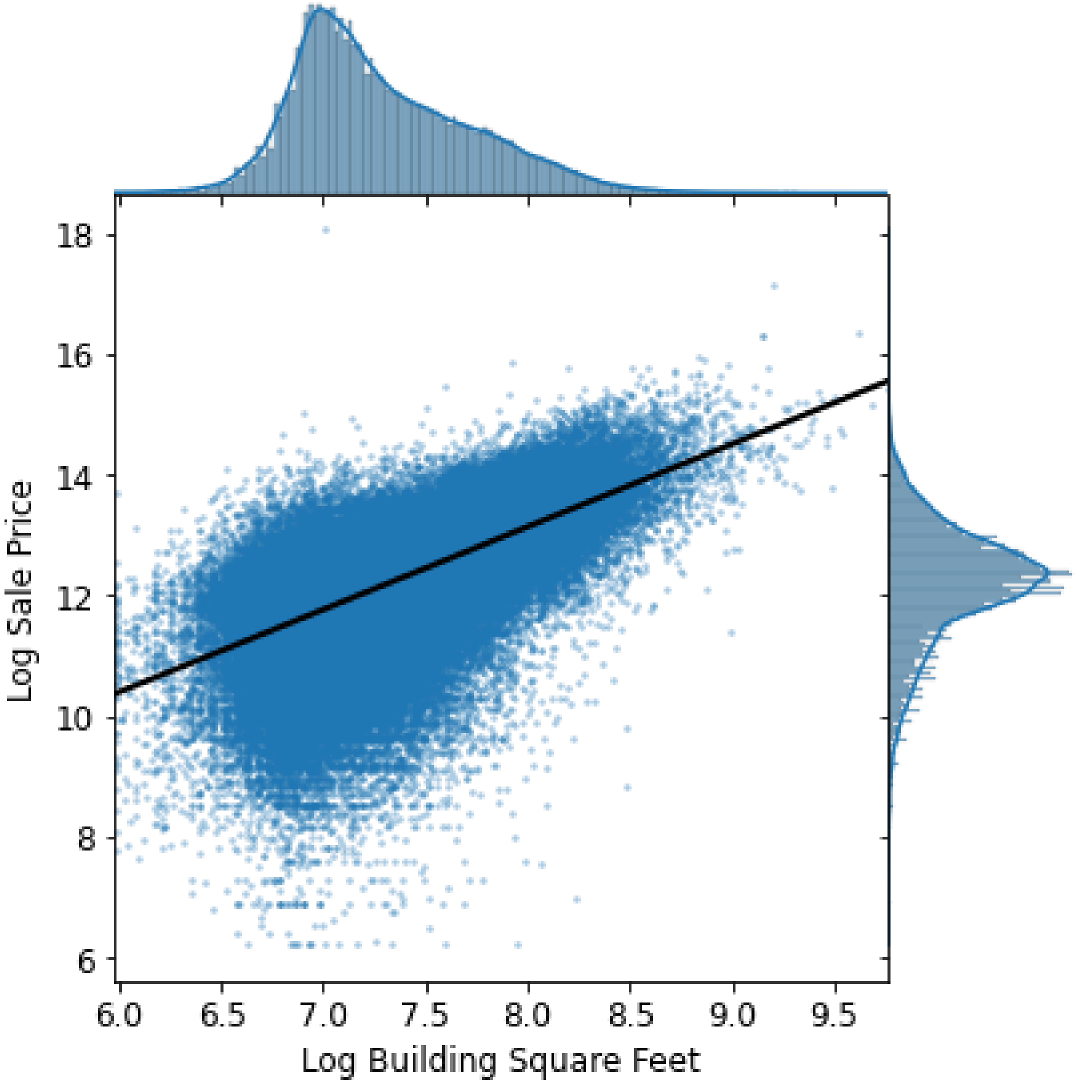
It shows the minimum sale price as $1, so I would filter for a minimum price of something like 1000.

1.6 Question 3c

In the visualization below, we created a `jointplot` with `Log Building Square Feet` on the x-axis, and `Log Sale Price` on the y-axis. In addition, we fit a simple linear regression line through the bivariate scatter plot in the middle.

Based on the following plot, would `Log Building Square Feet` make a good candidate as one of the features for our model? Why or why not?

Hint: To help answer this question, ask yourself: what kind of relationship does a "good" feature share with the target variable we aim to predict?



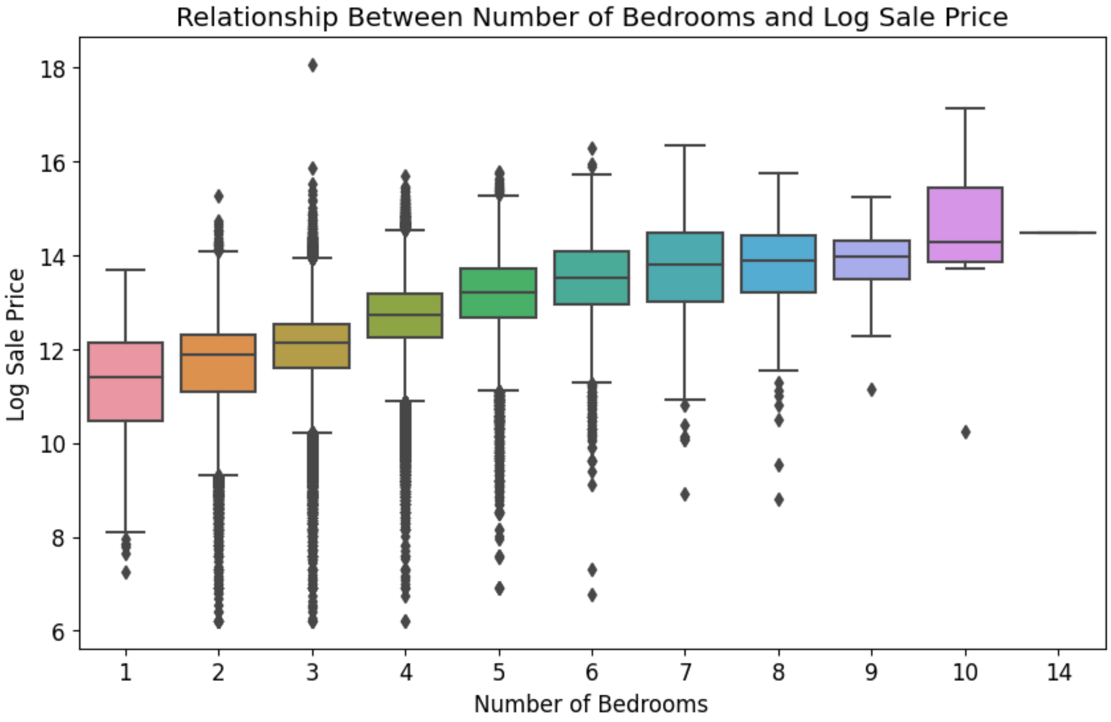
Yes it would make a good candidate. It has decent correlation, being somewhat tightly compacted around the line, and nice linearity when compared with Log Sale Price.

1.7 Question 5c

Create a visualization that clearly and succinctly shows if there exists an association between `Bedrooms` and `Log Sale Price`. A good visualization should satisfy the following requirements: - It should avoid overplotting. - It should have clearly labeled axes and a succinct title. - It should convey the strength of the correlation between `Sale Price` and the number of rooms: in other words, you should be able to look at the plot and describe the general relationship between `Log Sale Price` and `Bedrooms`

Hint: A direct scatter plot of the `Sale Price` against the number of rooms for all of the households in our training data would result in overplotting (since there are only a small discrete number of bedrooms) - so **don't use a scatter plot**.

```
In [117]: plt.figure(figsize=(10, 6))
          sns.boxplot(x='Bedrooms', y='Log Sale Price', data=training_data)
          plt.title('Relationship Between Number of Bedrooms and Log Sale Price')
          plt.xlabel('Number of Bedrooms')
          plt.ylabel('Log Sale Price')
          plt.show()
```


Relationship Between Number of Bedrooms and Log Sale Price
Log Sale Price
Number of Bedrooms