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?

The granularity of this data set is one row per property. So each row represents one property and its attributes/details

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.

Maybe to analyse property price trends such as what affects feature of the property affects its price significantly. The data could've also been collected to understand property within an area or for a site like zillow. The data also might have been from a government survey to understand properties in the area.

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. What is the impact of having a garage on the sale price of properties in Cook County? I would compare the distribution of Sale Price between properties with a garage (Garage Indicator column) and those without. To determine if there is a statistically significant difference, I can perform A-B testing on the two groups.
2. Does the noise from the airport (O'Hare Noise Indicator) affect the sale price of properties in Cook County? I would create a scatter plot of Sale Price vs O'Hare Noise Indicator to see if there is a correlation between the two variables.

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.

1. Is there a correlation between the Sale Price of a property and the race/ethinicity of the owner? For this we could compare the sale prices of the properties grouped by their race/ethnicity. Using TVD, we can determine if there is a significant difference in the sale prices of the properties.

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.

As the scale of the x-axis is determined by the most expensive house, it caused majority of the houses to be congregated at the far left. As the pricing of houses are not evenly distributed between the cheapest and the most expensive, and there are commonly more houses on the lower end of the scale, our graph is heavily skewed right. One way we can overcome this issue is by only visualising houses within the whiskers of the box-plot as that is where we can expect most of our house prices to be within. That is we should limit our visualisation to only show data within $[4.52 \cdot 10^4 - (1.5 \cdot (3.12 \cdot 10^5 - 4.52 \cdot 10^4)), 3.12 \cdot 10^5 + (1.5 \cdot (3.12 \cdot 10^5 - 4.52 \cdot 10^4))]$

This way we only get values up to the most extremes that are not outliers

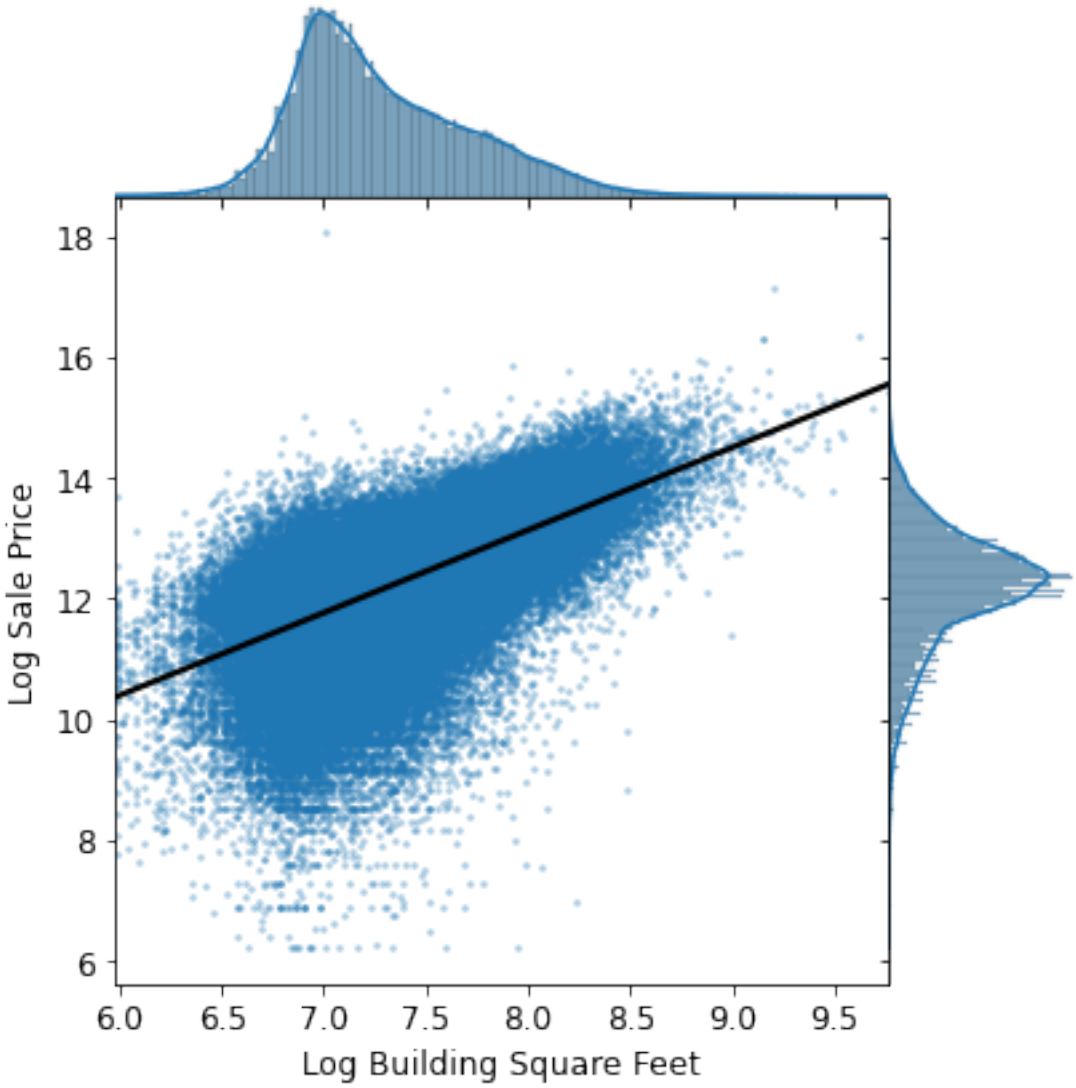
Alternatively, we could use a logarithmic scale to decrease the distance of extreme values from the rest of the data.

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?



From the plot, it seems there is a fairly strong positive linear correlation between the Log Building Square Feet and the Log Sale Price. There is a clear trend that as the Log Building Square Feet increases, the Log Sale Price also increases. There also does not appear to have any other underlying patterns in the plot. There also are not extreme outliers that would affect the linear regression line. Therefore, Log Building Square Feet would make a good candidate as one of the features for our model.

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