0.1 Question 1a

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

Each row represents the info of each house in Cook County, Illinois like its unique PIN, it's price, type of house, town code,... The granularity of this data set is very fine grain, because each row/entries shows the fine details of each house in Cook County, Illinois.

0.2 Question 1b

Why was this data collected? For what purposes? By whom?

You should watch Lecture 15 before attempting this question.

There was a descrepancies between housing taxation between higher income and lower income due to the so-called estimate market price. Specifically, people think that they are paying more housing tax for cheaper property because its value was overestimate, and the more expensive property owner are paying less tax because of the system that underestimate more expensiv houses and overestimate cheaper houses. This hurts the lower income class and benefits the higher income class, and they even use this to segregate by races by making it harder for a certain race to get a house in an area. Therefore this data was collected to aim to create a fair housing price prediction by the Cook County Assesor's Data Science office.

0.3 Question 1c

Craft at least two questions about housing in Cook County that can be answered with this dataset 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 datasets you would need to answer that question.

1) How does the presence of a garage (Garage Indicator) affect the sale price (Sale Price) of properties in Cook County? I would create a box plot or violin plot of Sale Price grouped by Garage Indicator (1 = garage present, 2 = no garage) to visualize the distribution of sale prices for properties with and without garages. Additionally, I would calculate the mean and median sale prices for each group to quantify the difference in sale prices.

2) What is the relationship between the age of a property (column Age) and its sale price (column Sale Price), and does this relationship vary by property class (col Property Class)? I would create a scatter plot of Age versus Sale Price, with points colored by Property Class to visualize the relationship between age and sale price across different property classes. I would also calculate the correlation coefficient between Age and Sale Price for each property class to quantify the strength and direction of the relationship. Additionally, I might use a regression model to predict Sale Price based on Age and Property Class, including an interaction term to test if the effect of age on sale price varies by property class.

0.4 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.

Question is: How does the annual income of the property owner (Annual Income) relate to the sale price (Sale Price) of the property, and does this relationship differ by the owner's race/ethnicity (Race/Ethnicity)?

I would create a scatter plot of Annual Income versus Sale Price, with points colored by Race/Ethnicity to visualize the relationship between income and sale price across different racial/ethnic groups. Additionally, I would calculate the correlation coefficient between Annual Income and Sale Price for each racial/ethnic group to quantify the strength and direction of the relationship. To further analyze the differences, I might use a multiple regression model with Sale Price as the dependent variable and Annual Income, Race/Ethnicity, and their interaction term as independent variables to test if the effect of income on sale price varies by race/ethnicity.

0.5 Question 1e

Look at `codebook.txt` to see some of the unique regional features CCAO utilizes, such as `O'Hare Noise`. Now imagine you were in charge of predicting the `Sale Price` of houses in **your hometown** (your actual real life hometown/city - not the data provided). Propose a feature that you would want to collect specific to your location and hypothesize why it might be useful in predicting the sale price of houses.

I would create a school district quality rating or a proximity to economic hubs.

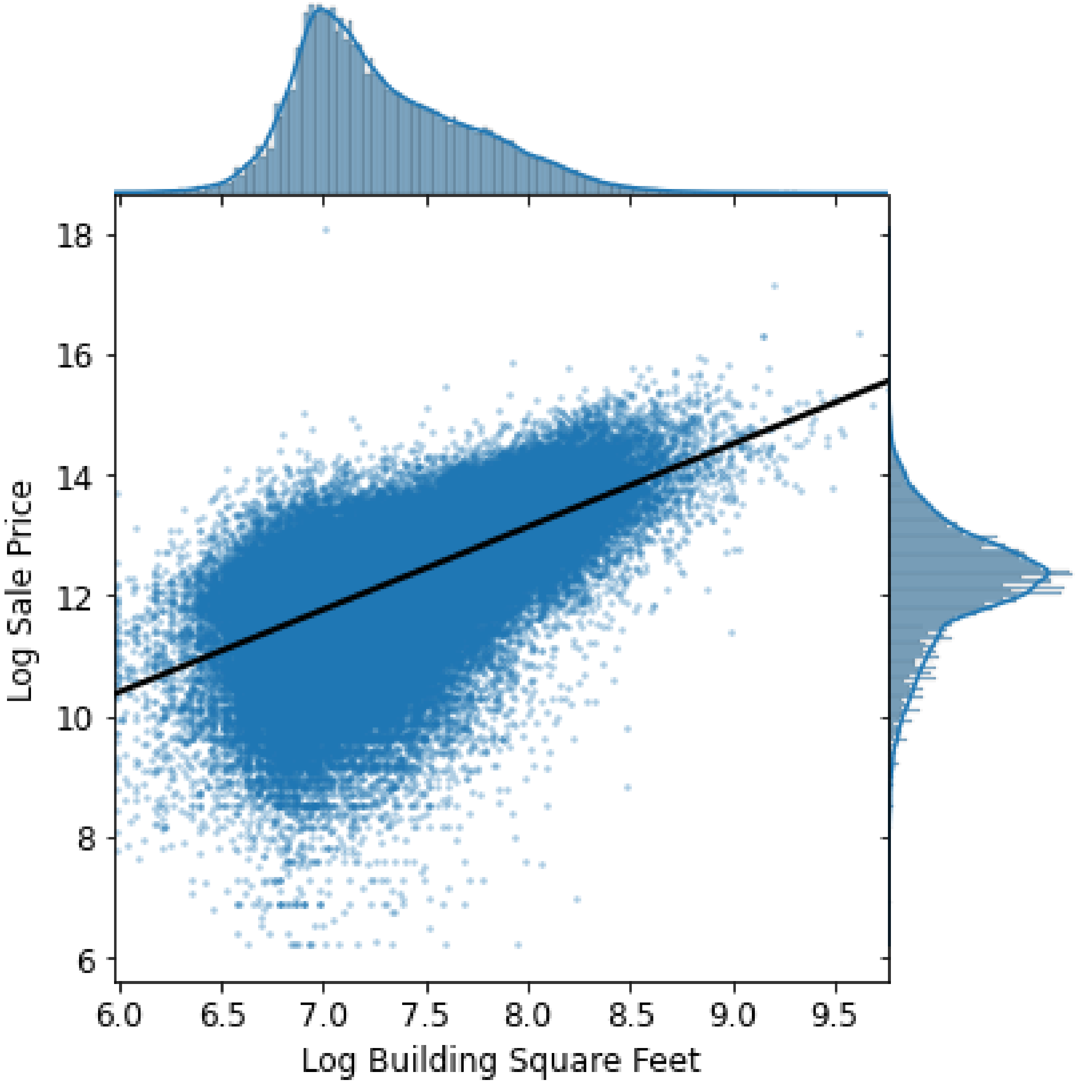
Proximity to major employers or economic hubs is often a significant factor in determining property values. Homes closer to these hubs tend to be more desirable because they reduce commute times and provide easier access to jobs, amenities, and services. I think this feature could help capture the economic vitality of different neighborhoods and predict higher sale prices for properties located near these hubs. Same with school district quality rating, properties located in higher-rated school districts will have higher sale prices compared to those in lower-rated districts. This relationship might be particularly strong in suburban areas where families with children are a significant portion of the home-buying population.

0.6 Question 3b

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?



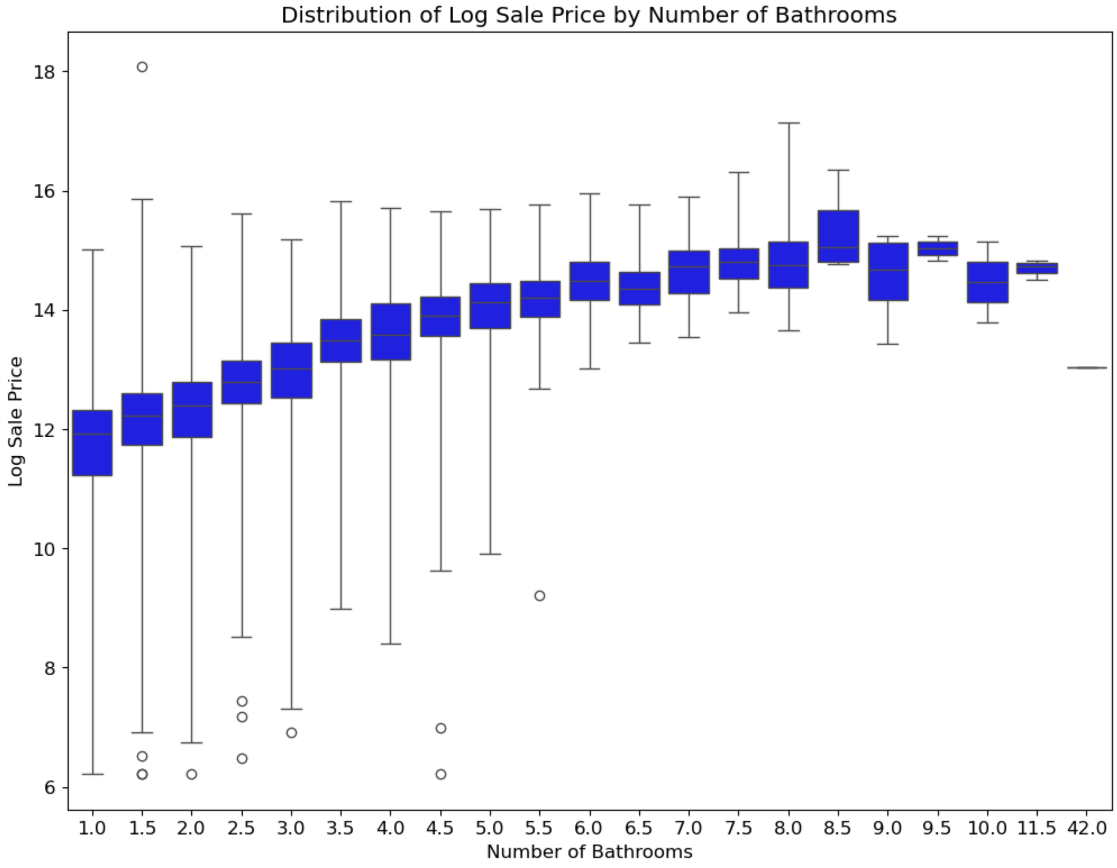
Based on the visualization, Log Building Square Feet seems like a good feature to include in the model. The scatter plot shows a clear upward trend, meaning that as the log of the building's square footage increases, the log of the sale price also tends to increase. The fitted regression line fits the data pretty well, and the points are clustered closely around it, which suggests a strong linear relationship. Since linear regression works best when there's a linear relationship between the features and the target variable, Log Building Square Feet seems like a solid choice. It helps explain how the size of a building affects its price, which makes it useful for predicting sale prices. So yeah, I'd say it's a good candidate for the model!

0.7 Question 5c

Create a visualization that clearly and succinctly shows if there exists an association between `Bathrooms` 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 `Bathrooms`.

Hint: A direct scatter plot of the `Sale Price` against the number of rooms for all of the households in our training data might risk overplotting.

```
In [135]: training_data['Log Sale Price'] = np.log(initial_data['Sale Price'])
          sns.boxplot(x='Bathrooms', y='Log Sale Price', data = training_data, whis = 6, color="blue")
          plt.title('Distribution of Log Sale Price by Number of Bathrooms')
          plt.xlabel('Number of Bathrooms')
          plt.ylabel('Log Sale Price')
          plt.show()
```


Distribution of Log Sale Price by Number of Bathrooms
Log Sale Price
Number of Bathrooms