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?

Each row represents various data for each house sold on the market. The granularity therefore would be something similar to the sale price of each house.

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.

The data was mostly likely collected to help either real estate housing agents or prospective buyers better understand the fluctuations of the housing market based on various aspects of the property.

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.

How do the distribution of Sale Prices vary across different ZIP codes in Cook County? I would create a box plot of ‘Sale Price’s grouped by ’ZIP Code’ to observe the variation in housing prices across different ZIP codes.

What is the trend in Sale Price over the years (Sale Year)? I would calculate the average ‘Sale Price’ by ‘Sale Year’ and plot a line graph to observe trends over time.

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.

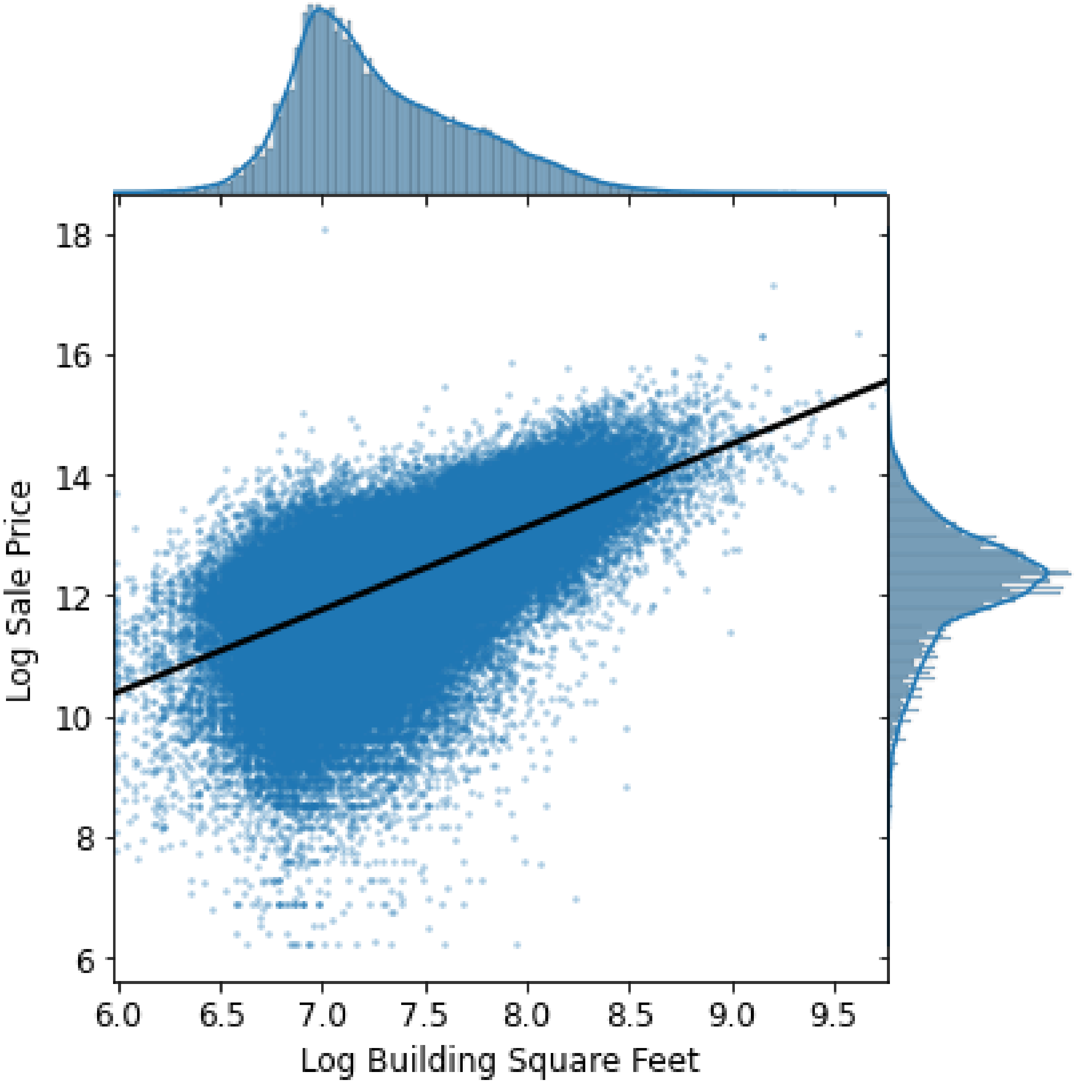
Is there a significant relationship between a homeowner's 'Race/Ethnicity' and the 'Neighborhood Code' of the property? I would use a stacked bar chart to show the distribution of Race/Ethnicity across neighborhoods.

1.5 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?



Log Building Square Feet is a decent candidate as a feature for our model as the scatter plot shows an upward-sloping trend, indicating that there's a positive correlation between Log Building Square Feet and Log Sale Price. This suggests that as the building square footage increases, the sale price tends to increase as well. The concentration of points around the line also indicates a strong relationship between the two variables. Also, the histograms show the distribution of Log Building Square Feet and Log Sale Price. Both distributions are reasonably concentrated, with Log Building Square Feet peaking around 7.5 and Log Sale Price peaking around 12, which supports the consistency and reliability of the relationship.

1.6 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

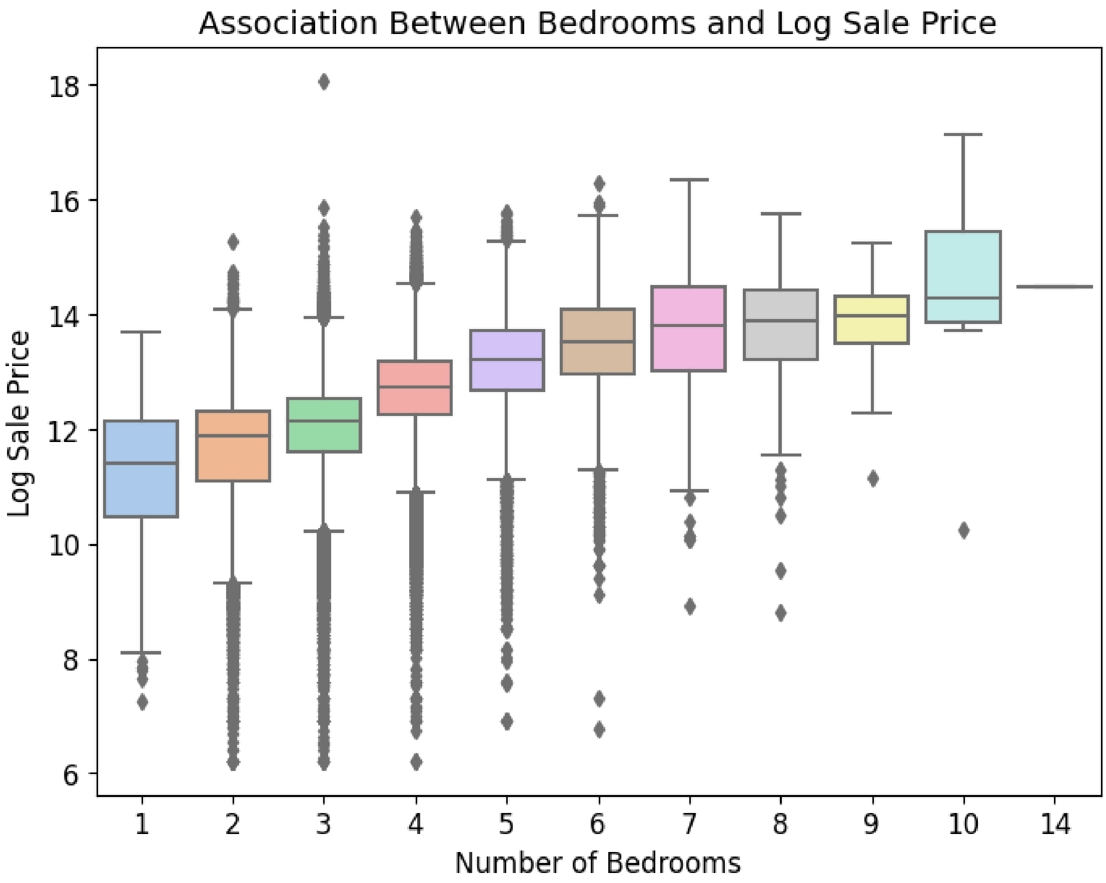
Hint: A direct scatter plot of the Sale Price against the number of rooms for all of the households in our training data would result in overplotting (since there are only a small discrete number of bedrooms) - so **don't use a scatter plot**.

```
In [141]: import seaborn as sns

          # Create a box plot
          plt.figure(figsize=(8, 6))
          sns.boxplot(
              x="Bedrooms",
              y="Log Sale Price",
              data=training_data,
              palette="pastel"
          )

          # Add labels and title
          plt.title("Association Between Bedrooms and Log Sale Price", fontsize=14)
          plt.xlabel("Number of Bedrooms", fontsize=12)
          plt.ylabel("Log Sale Price", fontsize=12)

          # Show the plot
          plt.show()
```


Association Between Bedrooms and Log Sale Price
Log Sale Price
18
16
14
12
10
8
6
1
2
3
4
5
6
7
8
9
10
14
Number of Bedrooms