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 1

1.1.1 Part 1

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 a single home in Cook County.

1.1.2 Part 2

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.

This data was possibly collected for the purpose of real estate. For example companies like Zillow, Redfin, or Remax may find this data useful to provide information to users/customers about homes in Cook County, Illinois.

1.1.3 Part 3

Certain variables in this data set contain information that either directly contains demographic information (data on people) or could when linked to other data sets. Identify at least one demographic-related variable and explain the nature of the demographic data it embeds.

Similarly, the columns 'Neighborhood Code', 'Town Code', 'Town and Neighborhood', 'Longitude', and 'Latitude' combined with other data sets could give away information about the demographics of the owners/renters of the home, this possibly includes their economic, racial, or gender demographic. Data sets that give away the relationship between locations and these demographic factors allows you to perform analysis to can find the likelyhood of a particular home owner being of a certain demographic.

1.1.4 Part 4

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.

I would create a scatterplot of 'Age' and 'Sale Price' to see if there is a correlation between older homes and lower prices by then finding the correlation coefficient. I wonder is the type of porch, frame or masonary, adds a greater value to the home. I would find this by first filtering out the 3 values in the 'Porch' column which indicate no porch. Then groupby the Porch column and aggregate by average to find the average 'Sale Price' for each type of porch.

1.2 Question 2

1.2.1 Part 1

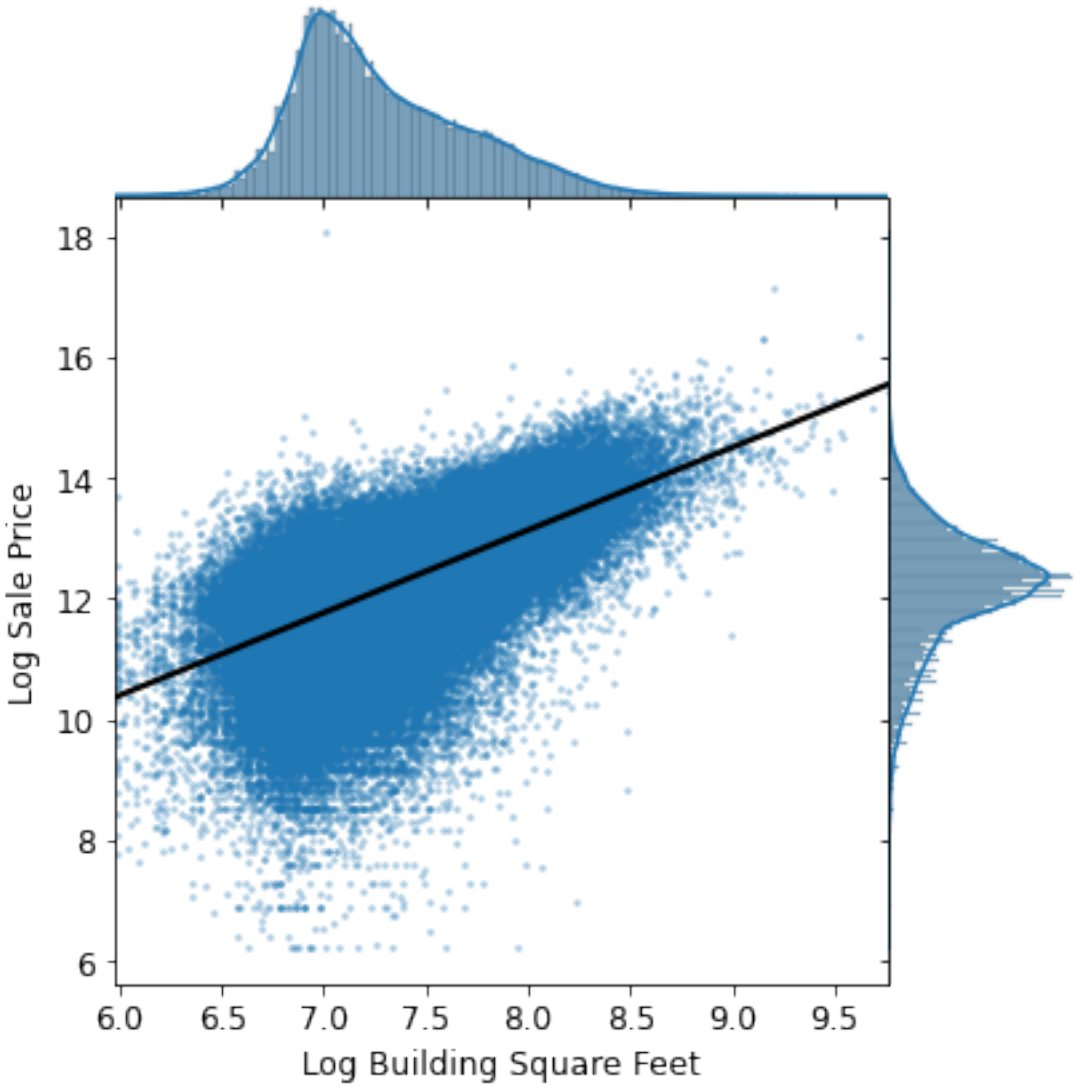
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards as the autograder may not work otherwise.

One issue with the visualization is that by including the outliers the scaling of the Sale Price gets unnecessarily extended causing the graphs to be difficult to read as the bulk of the values/distribution are squished to a small portion of the visualization. One way to overcome this is to log transform the Sale Price column of the training_data dataframe.

1.2.2 Part 3

As shown below, we created a joint plot 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, does there exist a correlation between `Log Sale Price` and `Log Building Square Feet`? Would `Log Building Square Feet` make a good candidate as one of the features for our model?



Yes based on the visualization there exists a correlation between Log Sale Price and Log Building Square

Feet. There is a positive correlation between the two as the fitted simple linear regression line is upward sloping. Additionally it seems that the simple linear regression line is fairly representative of the overall trend of the graph since it is not pulled by outlies which makes it fairly homoscedastic. Therefore, Log Building Square Feet would make a good candidate as one of the features for our model.

1.2.3 Part 3

Create a visualization that clearly and succintly 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 succinct title. - It should convey the strength of the correlation between the sale price and the number of rooms.

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 [ ]: import seaborn as sns
        ax = sns.violinplot(data=training_data, x="Bedrooms", y="Log Sale Price", saturation=0.5, palet
        ax.set_title("Distribution of Log Sale Price by Number of Bedrooms");
```

1.2.4 Part 3

It looks a lot better now than before, right? Based on the plot above, what can be said about the relationship between the houses' `Log Sale Price` and their neighborhoods?

There is no strong relationship between the houses' Log Sale Price and their neighborhoods. The median of each neighborhood code's log sale price are all rougly around 12.3 and the range are all rougly the same as well, except for the 120 neighborhood code.