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?

Each row represents a single household from Cook County, Illinois

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.

Housind data can be used to explore trends in housing market and property value. These data are essential for online real estate agents such as zillow or trulia to showcase households, compare property values and predict housing price based on their features such as the number of bedrooms, the lot area, the neighborhood and etc.

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.

Town and neighborhood - uniquely identify neighborhoods across townships. We can learn about population density, race, crime rates and the average commute time from this kind of data.

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.

Two questions: - what features determines the value of a house: I would create a scatter plot based on two features (Building Square Feet, Age) and find the pattern of sale prices. - what is the average age of a property in Cook County: I would calculate the summary statistic for sale price and look the mean.

1.2 Question 2

1.2.1 Part 1

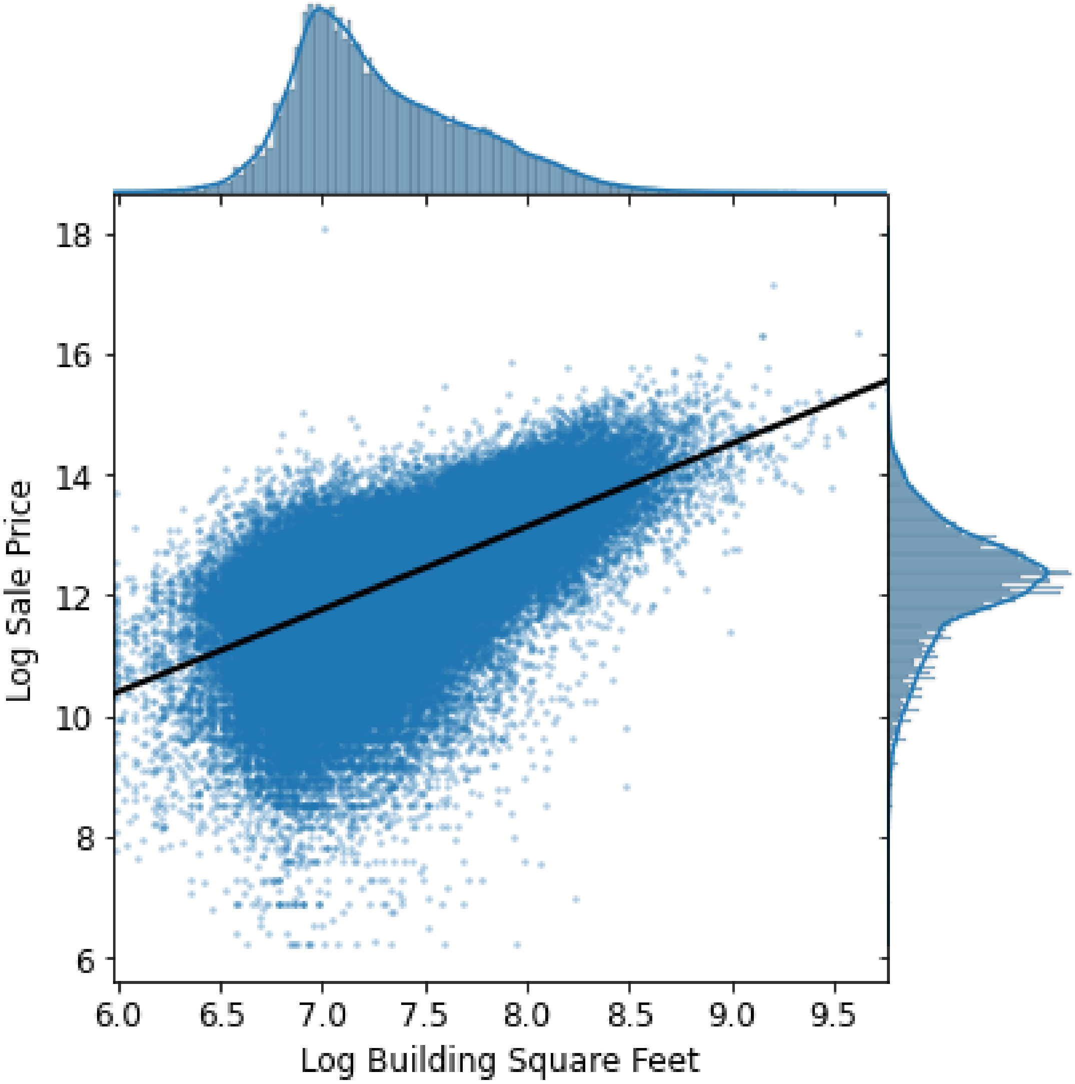
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

The distribution seems inappropriate due to the outlier in sale price. When running training_data[‘Sale Price’].describe(), the 75% of sale price is 3.12e+5 but there is a maximum 7.10e+7. We could filter out the outlier or we could transform our sale price data to a smaller scale (i.e. logarthmic).

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?



There exists a correlation between Log Sale Price and Log Building Square Feet. It makes a good candidate as one of the features for our model.

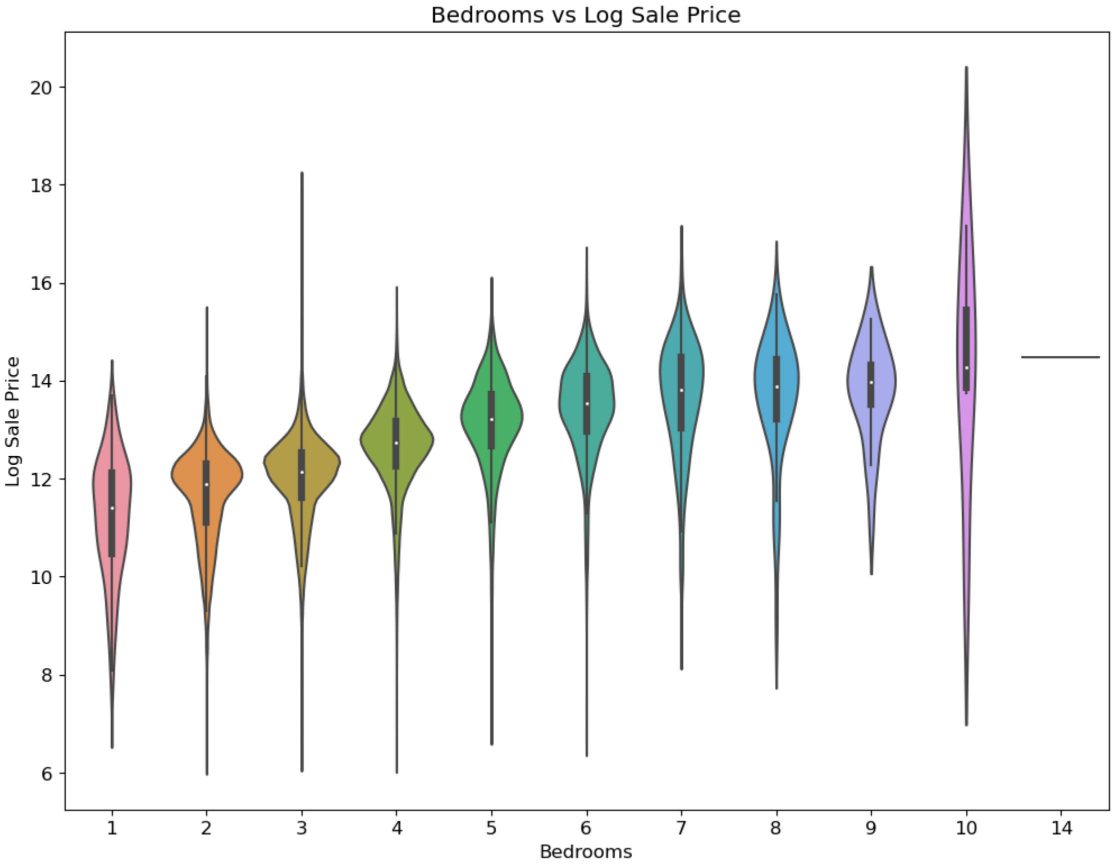
1.2.3 Part 3

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 [63]: sns.violinplot(data=training_data, x="Bedrooms", y="Log Sale Price")
         plt.xlabel("Bedrooms")
         plt.ylabel("Log Sale Price")
         plt.title("Bedrooms vs Log Sale Price")
```

```
Out[63]: Text(0.5, 1.0, 'Bedrooms vs Log Sale Price')
```



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?

Neighborhoods with the most number of houses tend to be cheaper than those with the less number of houses. We can see that Log Sale Price in neighborhoods with less number of houses are above the red margin line.