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 in the dataset represent detailed information and description about a property in each neighborhood 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 might be collected to be used to estimate property tax or determining the average market price of a house in each neighborhood in Cook County. Thus, the data might have been collected by a property developer or the government.

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.

Since we have a column that contains the sale price of the property, we can group it by each neighborhood and we can get a demographic related of the socioeconomic status of people live in the neighborhood. For example, if the average sale price of property is around 1 million dollar in the neighborhood can indicate that the neighborhood is more affluent with a higher income.

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.

1. I would create a line plot of "Sales Price" and "Sales Year" to visualize the trend of housing prices over time.
2. I would calculate the average for "Sales Price" and house "Age" in each neighborhood.

1.2 Question 2

1.2.1 Part 1

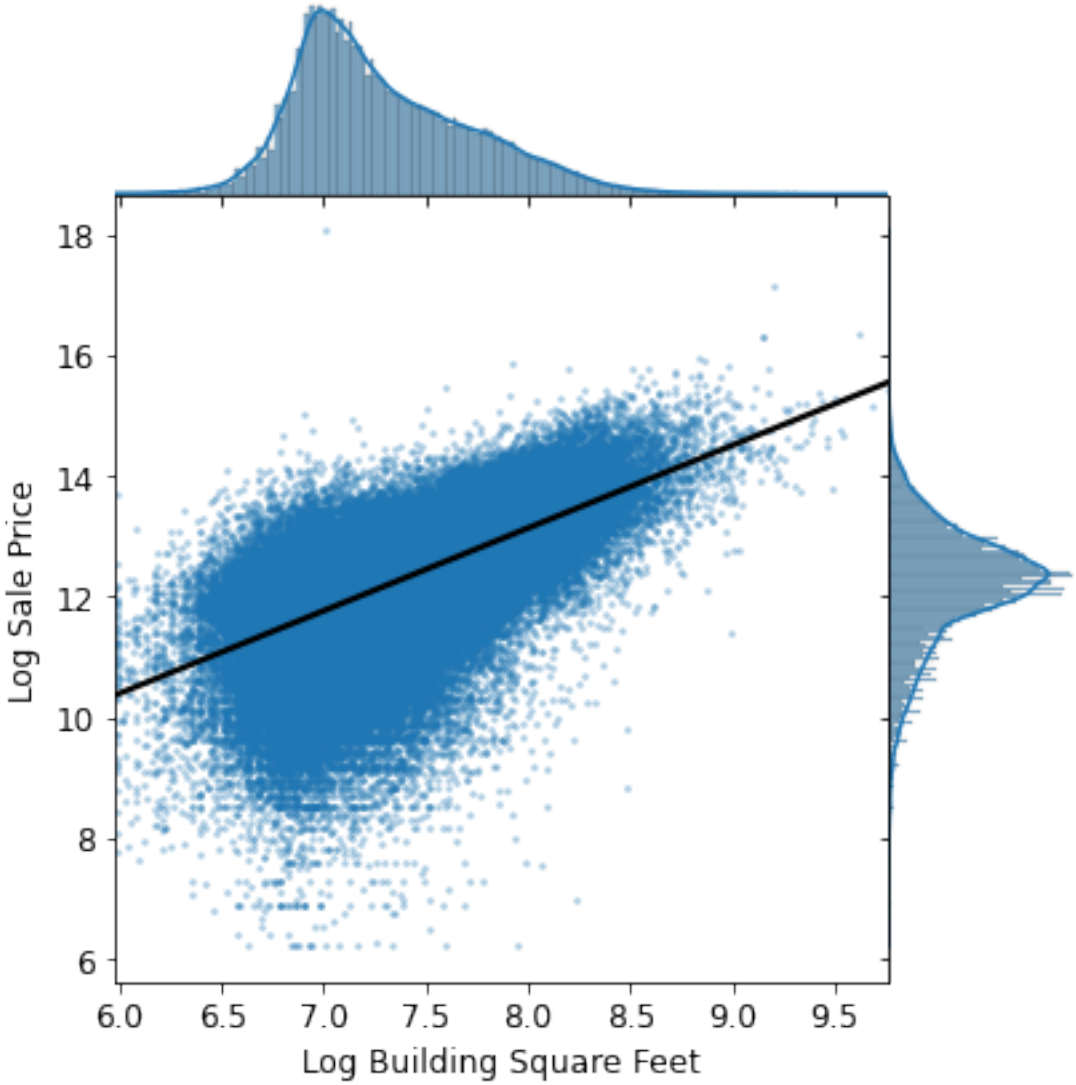
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

Our data range is too large/wide as our minimum value is 1 and our maximum is 71000000. The median of our data is 175000.0 which means that most of our data is at around this value, but due to outliers such as 71000000, our plot is too large and our distribution looks small.

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?



From our plot, we can see that our trend line is an increasing function which suggests that there's a positive

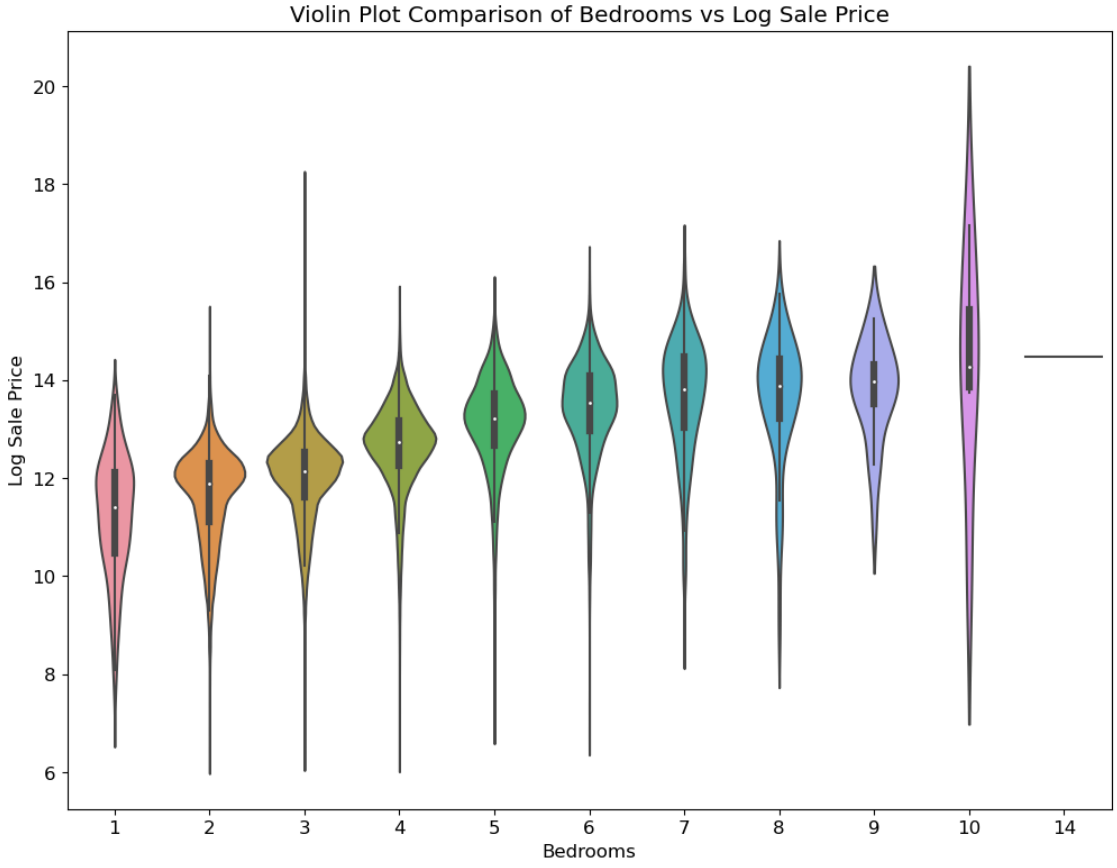
correlation between the two, thus the larger the square feet of a house will result in higher sale price.

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 [25]: sns.violinplot(data= training_data,x="Bedrooms", y="Log Sale Price")
         plt.title('Violin Plot Comparison of Bedrooms vs Log Sale Price');
         #We can see that from our plot that the larger the data, the median is higher, which signifies
         #rooms, there will be an increase in 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?

From the visualization above, we can see that most houses that top 20 neighborhoods have a mean of around 12 Log Sale Price and roughly have the same distribution. We can also see that houses in each neighborhood are sold at around the same price as most of the box plot has a narrow range of price except the neighborhood code of 120 which has a relatively wider range of price compared to the other neighborhood.