
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 information of a unique property in Cook County, Illinois. The granularity of this data is PIN (Unique Permanent Identification Number for each property).

0.2 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.

This data was collected by Cook County Assessment Office (because some of the variables in the dataset is assigned by Assessment Office), and was used to evaluate properties' values. This data could be used to understand the trend of properties' values over time.

0.3 Question 1c

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

"Neighborhood Code" : It doesn't directly reveal demographic information; however, the a specific neighborhood can reveal certain demographic information such as % race/ethnicity, age group, etc. of the residents. "Description": Because this variable contains the address of the property, it can be linked to other dataset to get more information about the household members.

0.4 Question 1d

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.

Question 1: What does the distribution of property classes (Residential Land, Commercial Land, etc) look like in Cook County? In order to achieve the answer, I would create a bar chart of categorical data in column "Property Class." Question 2: What is the income of the properties' owners? I would extract the address in column "Description" and link this information to other dataset that has information of the owners, then create a box plot of the income distribution to find the statistical summary of this data. Question 3: How does the distribution of the Sale Price of properties in Cook County look like? I would create a histogram from column "Sale Price" in this dataset to find out.

0.5 Question 2a

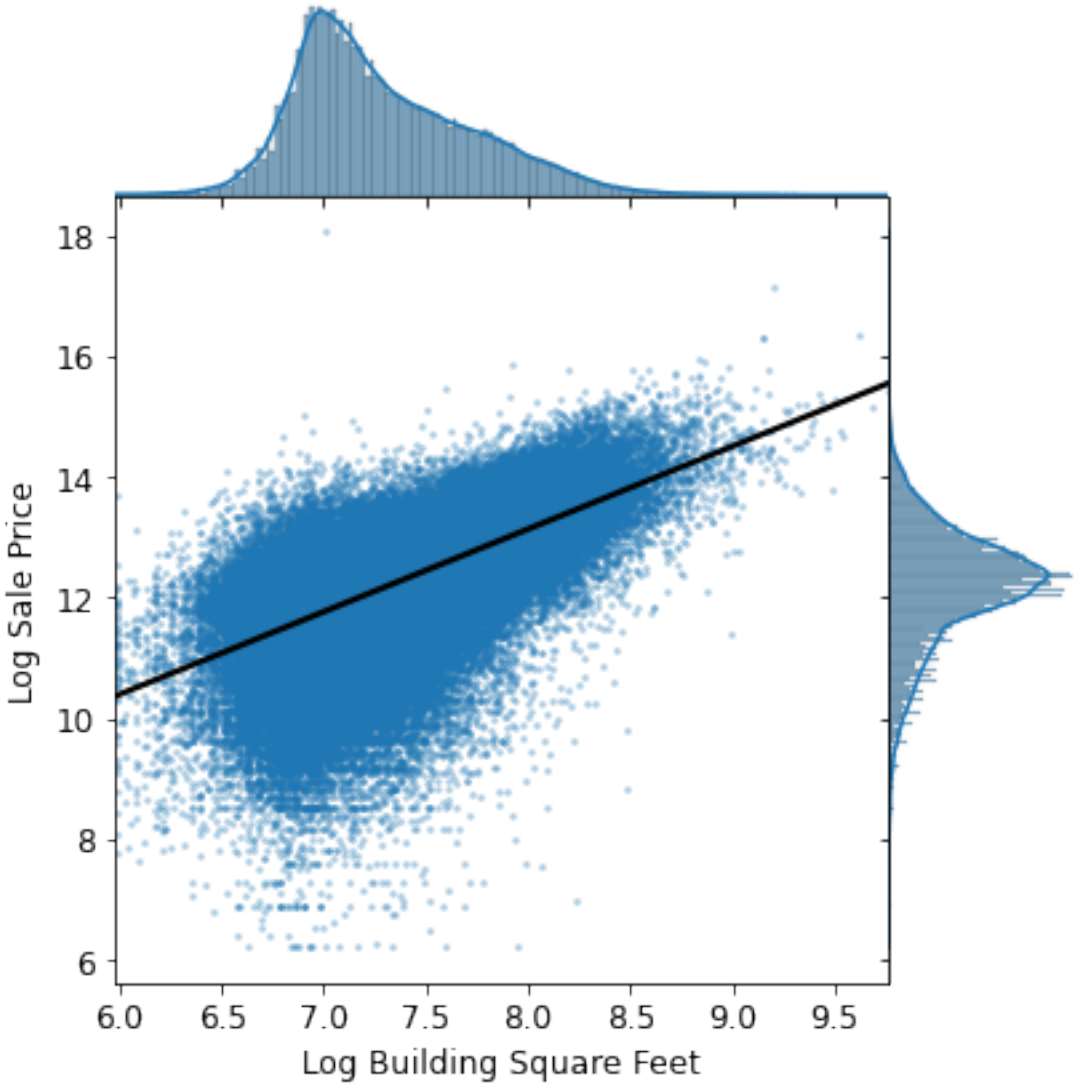
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

One issue with the visualization above is the Sale Price takes on large value (in 10,000) while some of them might have unusually small value (min = 1). This will squeeze our distribution to the point that we can not see clearly how it is distributed. One way we can overcome is to scale down the Sale Price values by taking the log. Another way of is to remove unusual small values (e.g.: 1) because these values do not have meaningful representation of real-life house price, and they wouldn't impact our distribution.

0.6 Question 3c

As shown 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?



I think `Logf Building Square Feet` would make a good candidate as one of the features for our model

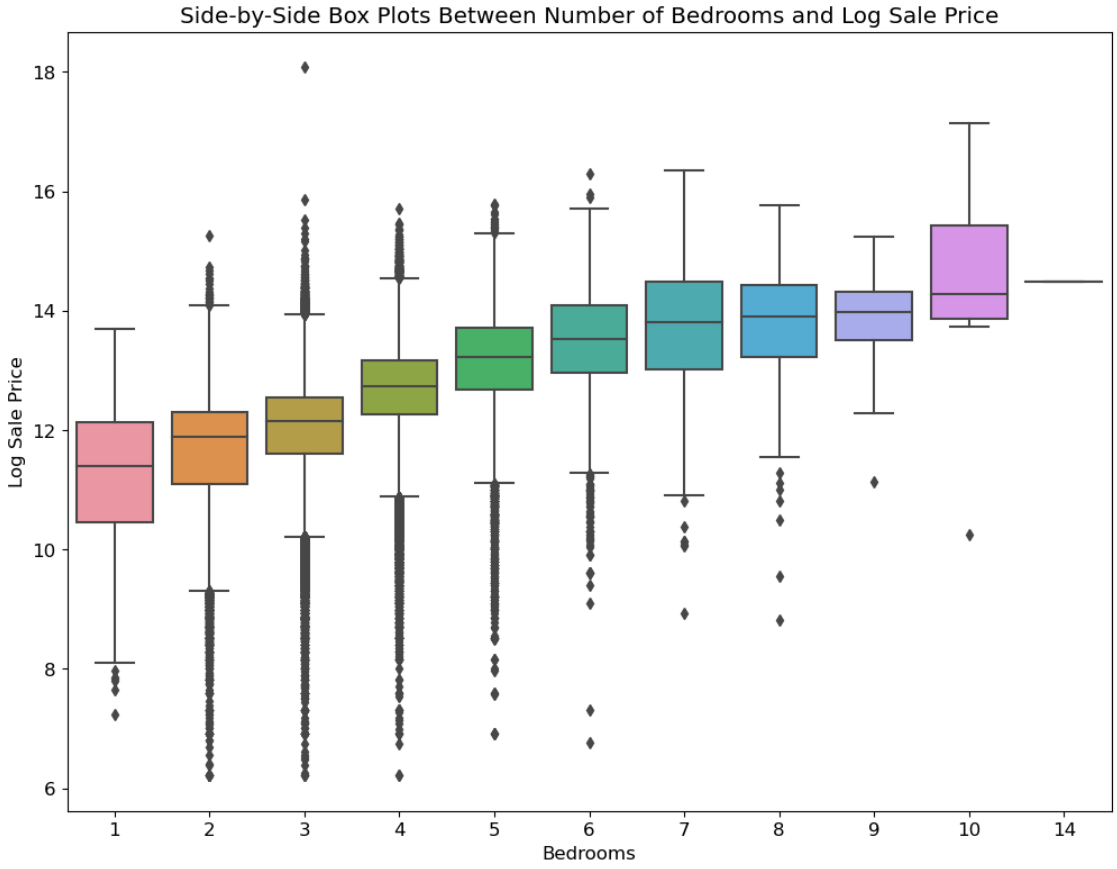
because these two variables have a positive association. While there is a large variability over lower value of Log Building Square Feet, most of the data lie around the line.

0.7 Question 5c

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 [117]: sns.boxplot(data = training_data, x = 'Bedrooms', y = 'Log Sale Price')
          plt.title("Side-by-Side Box Plots Between Number of Bedrooms and Log Sale Price");
```



0.8 Question 6c

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? Is there a relationship?

Based on the plot above, it seems that the top 20 neighborhoods have roughly the same median of log sale prices and this is also roughly the same as the median log sale price of the whole county. Additionally, those neighborhood only differs in their variability of the price distribution, but it could be dependant on how many properties are there in the neighborhood. With this being said, there appears a constant relationship between houses' Log Sale Price and their neighborhoods; in other words, we can use the a constant, in this case it is the median sale price of each neighborhood, to predict the sale price of the whole county and vice versa.