
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?

I think each row represents a housing property. It can be a house, condo, apartment complex, etc. detailing its location, land size, size of property, and all the amenities.

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 dada seems to have a lot of details about housing properties. Things like square footage, location, noise polution, and its last sale price. To me this seems like something a realtor would want to have to know the details of all the houses on the market.

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.

I would say that the deed number would be one of these variables. Someone could use this deed number and find it on another dataset and find the owners location and other personal data.

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.

I would want to see how many properties have catherdral ceilings in the county in each town. I would create a barplot to see the number of cathedral ceilings and town code. I would also want to see where there are the properties with the highest land square feet. I would create a scatter plot of Longitude and latitude and have them grouped by town.

0.5 Question 2a

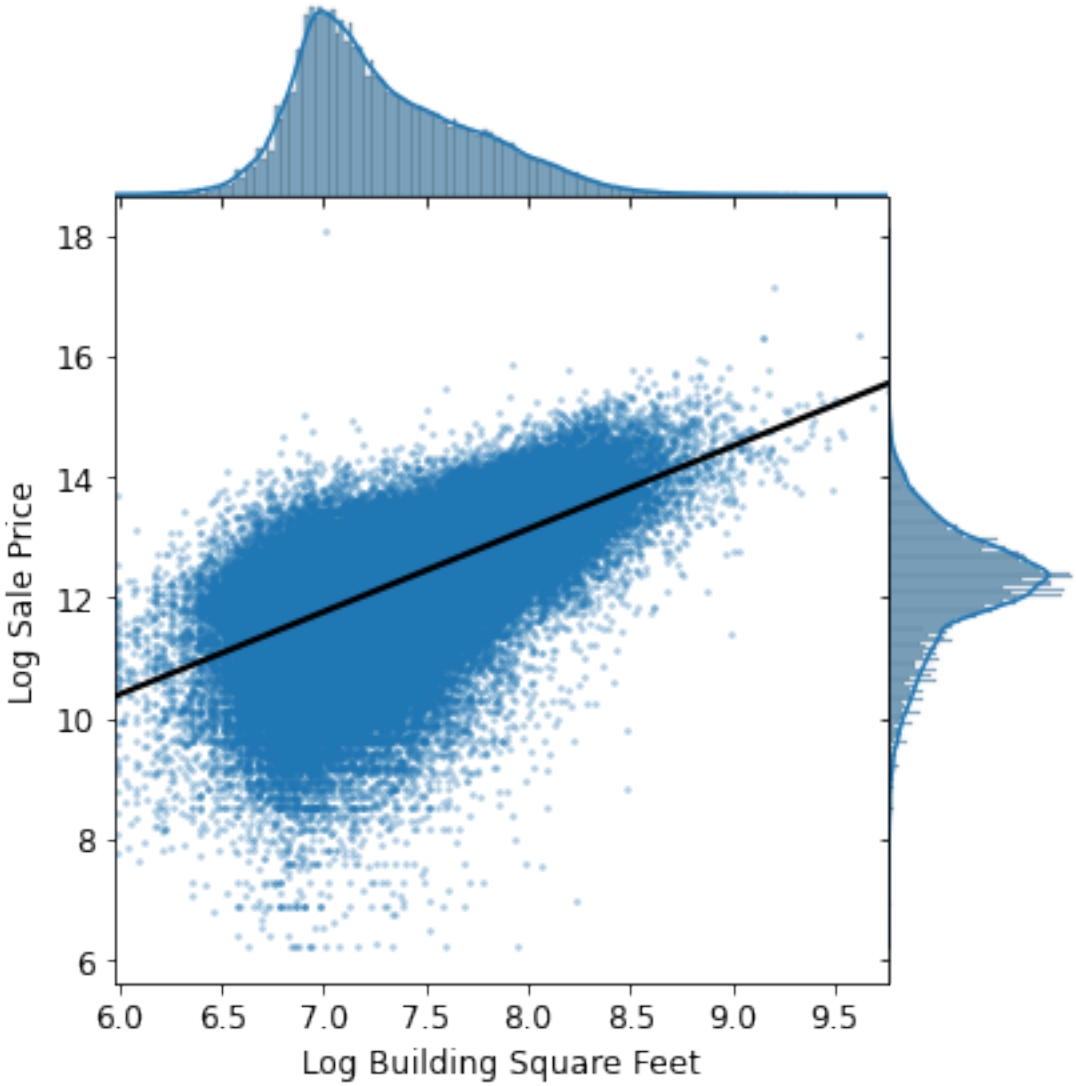
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

It seems like there are some outliers that really far out. There sale price is super high compared to the rest. Once way could be to get rid of it, but I don't think that is the best way to do this. I think the best way to do fix this is to take the log or square root of the sale price column and plot using that.

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?



This does seem like a good candidate as one of the feastures we can use. It seems like there is a positive

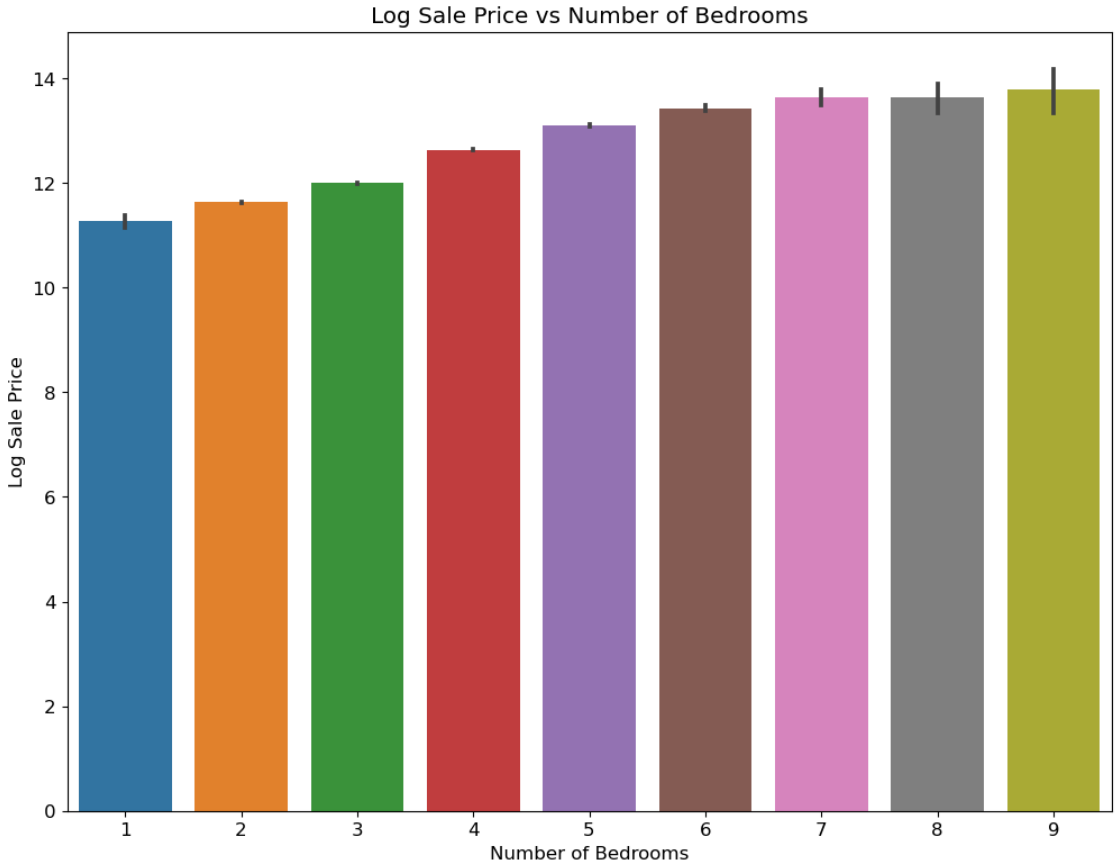
corelation between these two variables.

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 [25]: sns.barplot(data= training_data, x= 'Bedrooms',y= 'Log Sale Price')
         plt.xlabel('Number of Bedrooms')
         plt.title('Log Sale Price vs Number of Bedrooms')
         plt.show()
```



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?

I think there is a realtionship between these two variables. It seems like the median for all these neghborhoods is around 12. This makes me think that these houses are all similiar since the median for all the neighborhoods are around the same value.