
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?

Based on the columns in the dataset and the values they contain, each row in this dataset represents information about a specific real estate property. The dataset appears to be a collection of features and attributes related to individual properties. These features include information such as property characteristics, location, sale details, and physical attributes. The dataset provides a comprehensive view of different aspects of each property, and each row likely corresponds to a unique property or real estate transaction.

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.

Government agencies might collect this data for urban planning and infrastructure development. Understanding property characteristics and their locations can help in making decisions related to zoning, transportation, and urban development. **Property tax authorities** may use this data to assess property taxes based on the property's characteristics, size, and location.

0.3 Question 1c

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: I would like to understand the distribution of property prices in Cook County. To answer this question, I would create a histogram of 'Sale Price' values, with price ranges on the x-axis and the number of properties falling into each range on the y-axis. This would provide insights into the range and frequency of property prices within the dataset. Additional datasets are not required for this question.

Questi 2on: I'm interested in examining the relationship between property age and sale prices in Cook County. To answer this question, I would calculate the mean sale price for properties in different 'Age Decade' categories. I'd use a bar chart, with 'Age Decade' on the x-axis and the mean sale price on the y-axis. This would help identify whether older or newer properties tend to have higher or lower sale prices. No additional datasets are needed for this analysis.

0.4 Question 1d

Suppose now, in addition to the information already contained in the dataset, you also have access to several new columns containing demographic data about the owner, including race/ethnicity, gender, age, annual income, and occupation. Provide one new question about housing in Cook County that can be answered using at least one column of demographic data and at least one column of existing data and provide the type of analytical tool you would use to answer it.

Question: Is there a relationship between the race/ethnicity of property owners and the age of properties they own in Cook County? To answer this question, we can use scikit-learn to perform a correlation analysis. We will employ the 'Race/Ethnicity' column as a categorical variable and the 'Age' column as a continuous variable. By using a suitable tool for categorical vs. continuous data correlation, such as point-biserial correlation, we can assess whether there is a statistically significant relationship between the race/ethnicity of property owners and the age of the properties they own.

0.5 Question 2a

Using the plots above and the descriptive statistics from `training_data['Sale Price'].describe()` in the cells above, identify one issue with the visualization above and briefly describe one way to overcome it.

One issue with the visualization is that it does not effectively represent the full range of sale prices, especially the presence of extreme outliers with very high values. The histogram and box plot are heavily skewed to the right due to the existence of these outliers, making it challenging to discern the distribution for the majority of the data.

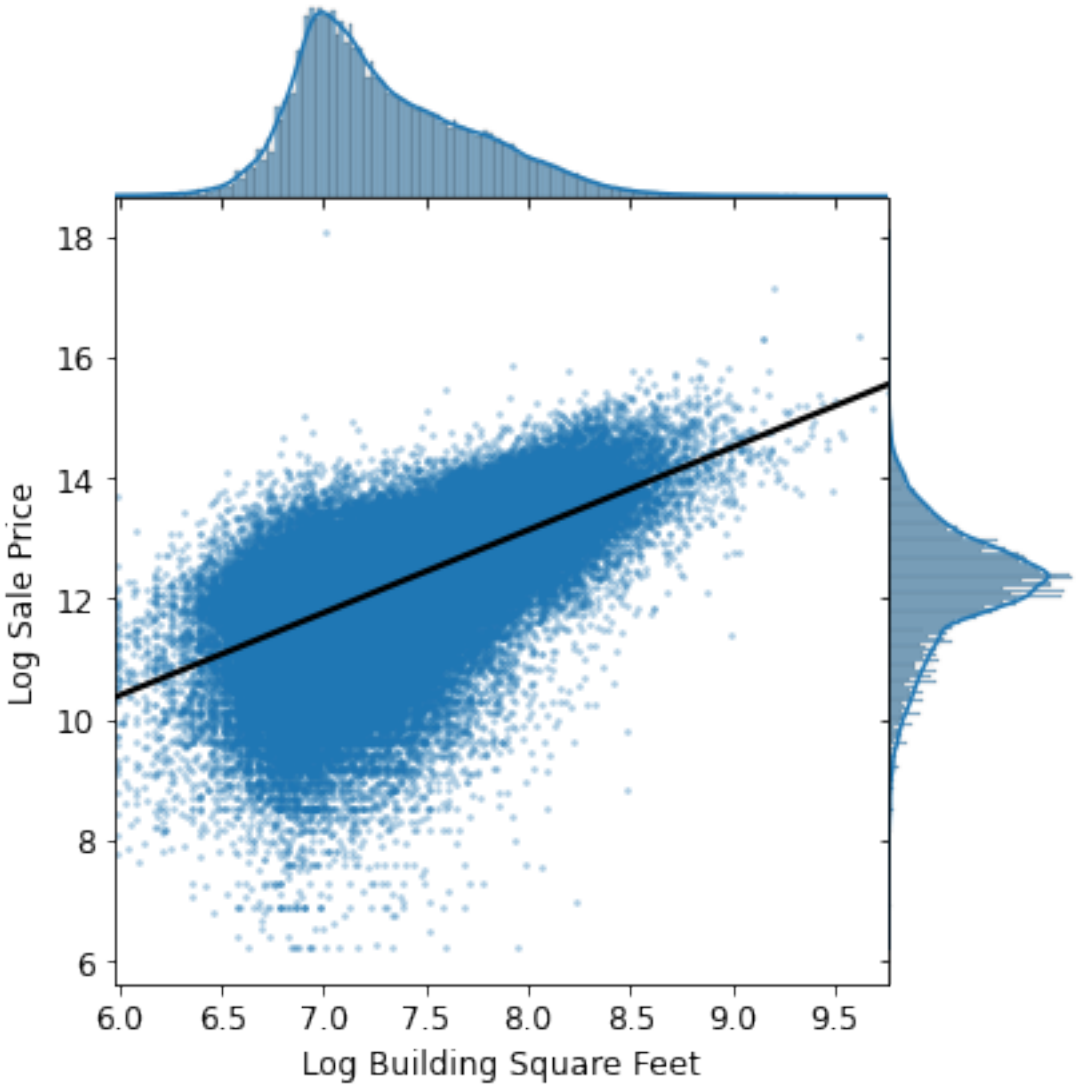
To overcom-ue, one way is to use a logarithmic scale for the y-axis of the histogram. By applying a logarithmic transformation to the sale prices, the distribution may appear less skewed, and the lower range of values can be more clearly visualized. This can help in gaining a better understanding of the central tendency and spread of sale prices while reducing the impact of extreme values on the visualization. Additionally, considering alternative visualizations such as a density plot or quantile-quantile (Q-Q) plot may also provide insights into the distribution and address the issue of outliers.

0.6 Question 3c

In the visualization 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?

Hint: To help answer this question, ask yourself: what kind of relationship does a "good" feature share with the target variable we aim to predict?



“Log Building Square Feet” appears to be a reasonable candidate as a feature for a predictive model. Firstly, on: The presence of an upward-sloping linear regression line in the scatter plot indicates a positive correlation between “Log Building Square Feet” and “Log Sale Price.” As “Log Building Square Feet” increases, “Log Sale Price” tends to increase, which is generally a desirable characteristic for a feat uAnd additionally, ship: The linear relationship between the two variables suggests that they can be modeled using a simple linear regression, which is a common and interpremodel.roach.

However, it’s essential to note the observations about the lower va” “Log Building Square Feet” “,Variance: The scatter plot’s”bulging left and upare Feet” suggests that the variance of “Log Sale Price” increases as “Log Building Square Feet” decreases. This non-constant variance might indicate heteroscedasticity, which could affect the model’s as sprovements: While “Log Building Square Feet” appears to be a promising feature, other factors might also influence “Log Sale Price.” It may be beneficial to consider interactions or additional features to capture the variation in the lower values of “Log Building Square Feet” and address the issue of non-constant variance.

0.7 Question 5c

Create a visualization that clearly and succinctly 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 a succinct title. - It should convey the strength of the correlation between `Sale Price` and the number of rooms: in other words, you should be able to look at the plot and describe the general relationship between `Log Sale Price` and `Bedrooms`

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]:
```
# Create a violin plot
plt.figure(figsize=(10, 6))
sns.violinplot(data=training_data, x='Bedrooms', y='Log Sale Price', palette='viridis')

# Set labels and title
plt.xlabel('Bedrooms')
plt.ylabel('Log Sale Price')
plt.title('Association between Bedrooms and Log Sale Price')
```

Out[25]:
```
Text(0.5, 1.0, 'Association between Bedrooms and Log Sale Price')
```