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 1a

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?

High level granularity

1.2 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.

It seems that the data was collected by the government or real estate companys because the data has attributes about houses and buildings which can be used for trend analysis, such as house price difference from area, building type and features of buildings.

1.3 Question 1c

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. Is there any difference on house price between areas?
    - I would create multiple scatter plots of sales price and neighborhood code
2. Does heating system type effects sales price?
    - Sorting data with Land square feet then separating data by Land square feet ranges
    - I would create scatter plots of sales price and heating system type

1.4 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.

Is there any community which is formed by race/ethnity per neighborhood?

I would create a histogram plot of race/ethnicity and neighborhood code.

1.5 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.

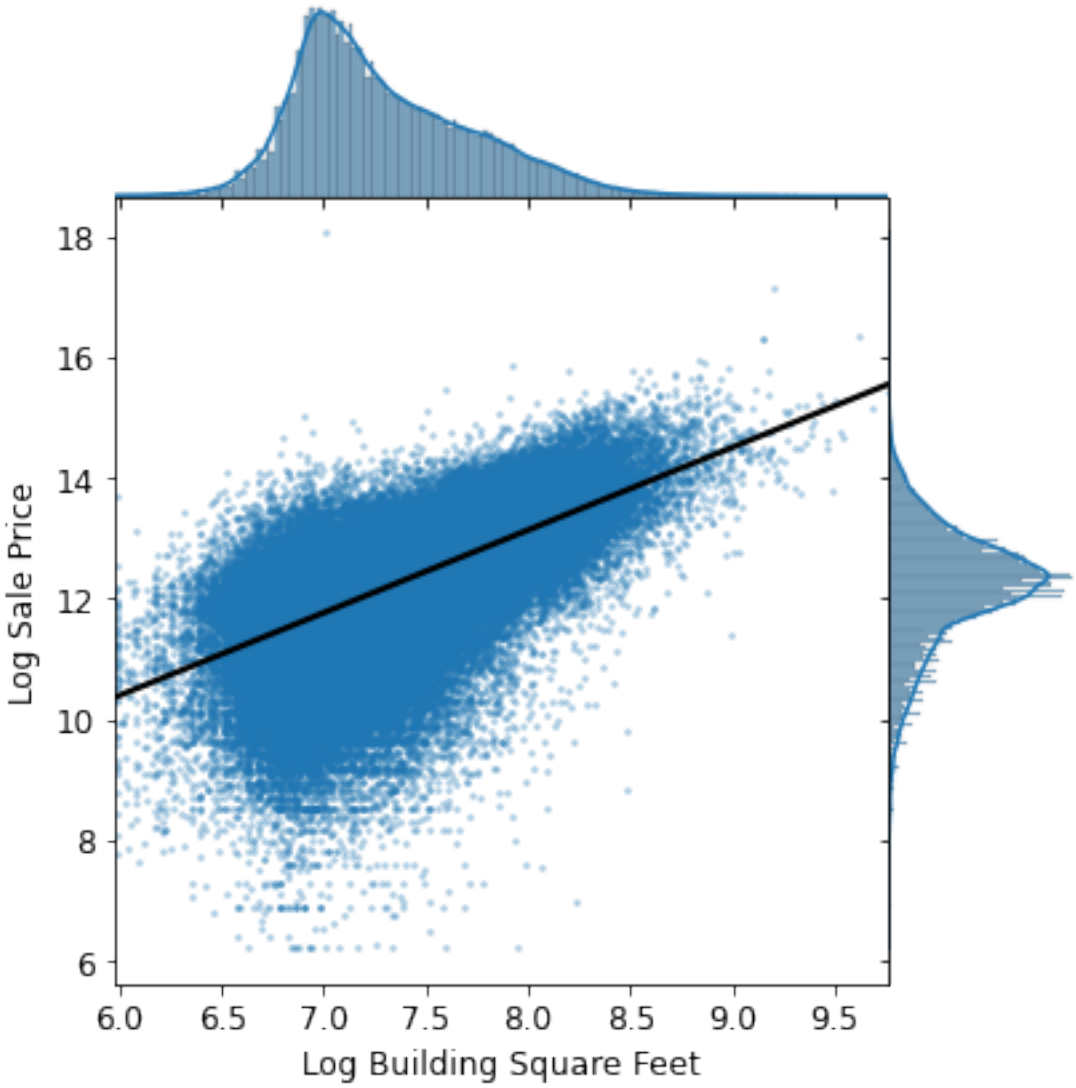
It is very difficult to read the plots since the value of x-axis is too large. Once the values of x-axis are replaced with a proper range, the plots can be readable.

1.6 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?



Yes, because the bivariate scatter plot seems to have positive correlation, which indicates there is a relationship, in other words, the feature is predictive.

1.7 Question 5c

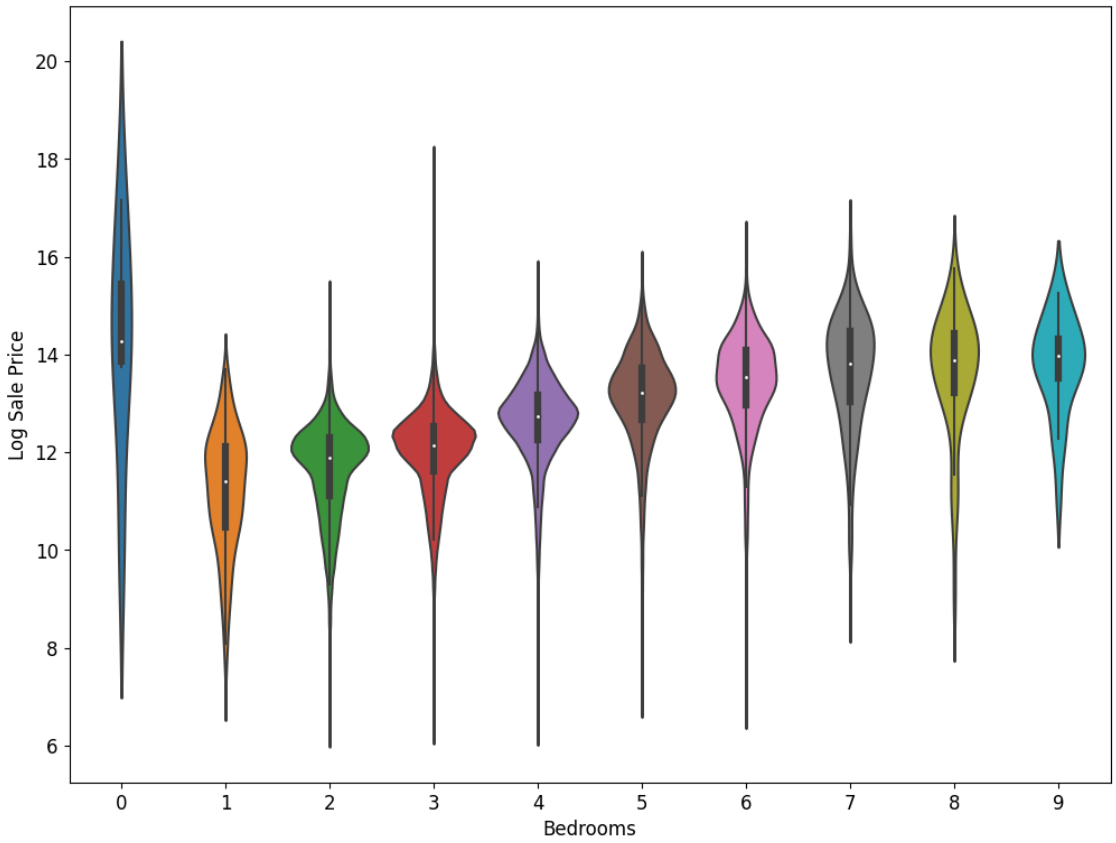
Create a visualization that clearly and succinctly shows if there exists an association between `Bedrooms` and `Log Sale Price`. A good visualization should satisfy the following requirements: - It should avoid overplotting. - It should have clearly labeled axes and a succinct title. - It should convey 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 would result in overplotting (since there are only a small discrete number of bedrooms) - so **don't use a scatter plot**.

In [30]:
```
br_logSP = training_data[["Bedrooms","Log Sale Price"]]
sns.violinplot(data=br_logSP, x='Bedrooms', y='Log Sale Price',
               title='Violin Plot of Log Sale Price by Num of Bedrooms')

#grouped_by_bedrooms = br_logSP.groupby('Bedrooms').mean()
#
#sns.jointplot(
#    x=grouped_by_bedrooms.index,
#    y=grouped_by_bedrooms['Log Sale Price'],
#    data=grouped_by_bedrooms,
#    #stat_func=None,
#    kind="reg",
#    ratio=4,
#    space=0,
#    scatter_kws={
#        's': 3,
#        'alpha': 0.25
#    },
#    line_kws={
#        'color': 'black'
#    }
#)
```

Out[30]:
```
<Axes: xlabel='Bedrooms', ylabel='Log Sale Price'>
```


Log Sale Price
20
18
16
14
12
10
8
6
0
1
2
3
4
5
6
7
8
9
Bedrooms