
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 is a single property within Cook County.

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.

The data was probably collected by government employees of Cook County, principally from the Assessor's office. The Assessor's office of Cook county was probably collecting this data to categorize and record general statistics about all of the properties within the county. This information could be useful to gain a relative understanding of the wealth and well-being of the county. It could potentially be used when the government wants to sell or lease a property to a company, and thus they have reasonable statistics to inform them how much to charge. Additionally, these statistics regarding property value can also allow the government to properly tax businesses and homeowers for property tax.

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.

What are the distributions of sale-prices for varying class property types?

- I would create a violin plot of the sale-prices against the property type. In order to do this, I would utilize the 'Property Class' and 'Sale Price' columns of the dataset, and feed those into sns.violin_plot as 'X' and 'Y' values.

What is the correlation between lot size and sale price?

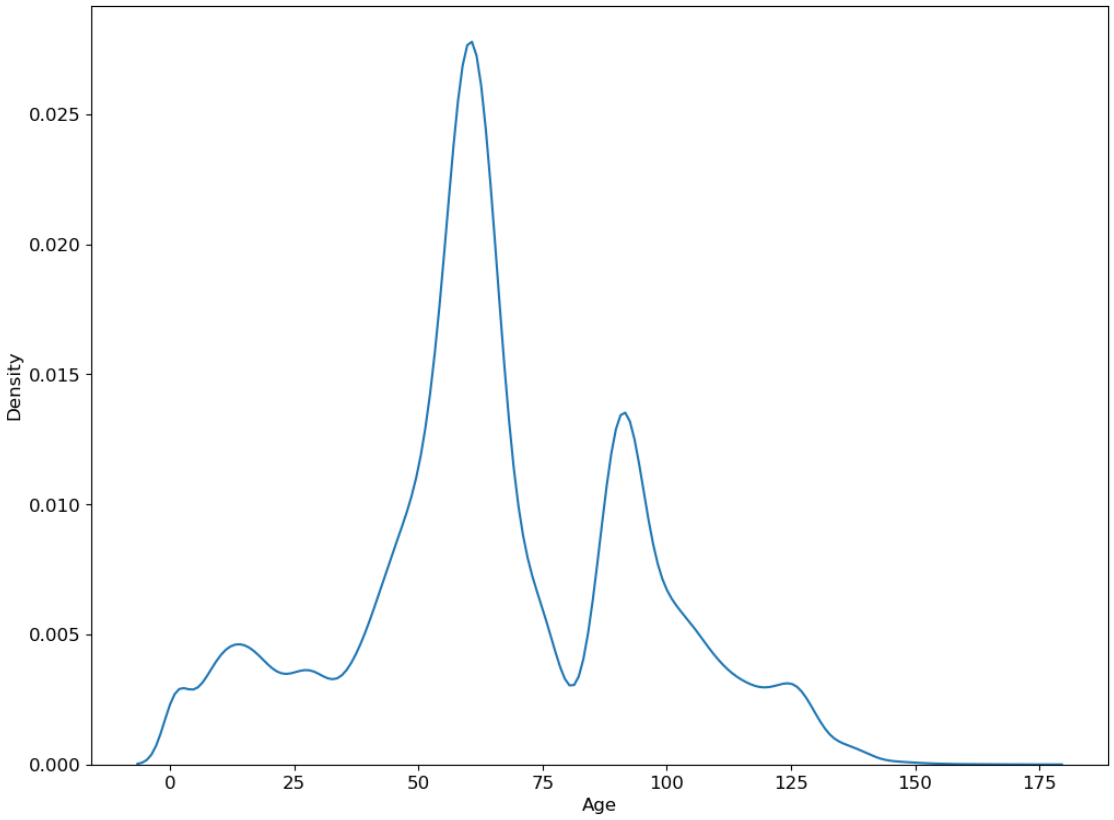
- I would create a scatter plot with lot size on the x-axis and sale-price on the y-axis. In order to do this, I would utilize the 'Lot Size' and 'Sale Price' columns of the dataset, and feed those into sns.scatterplot

What is the distribution of property ages within Cook County?

- I would create a KDE Plot of ages of properties. To do this, I would need utilize the 'Age' column of the dataset, and feed into sns.kdeplot. I could also create a Histogram of ages of properties, and would use sns.histplot instead. I have plotted these two graphs below. For histogram, I added an additional visualization variable of hue being the Property Class type, to also further visualize the if Property Class has any trends in terms of age distribution.

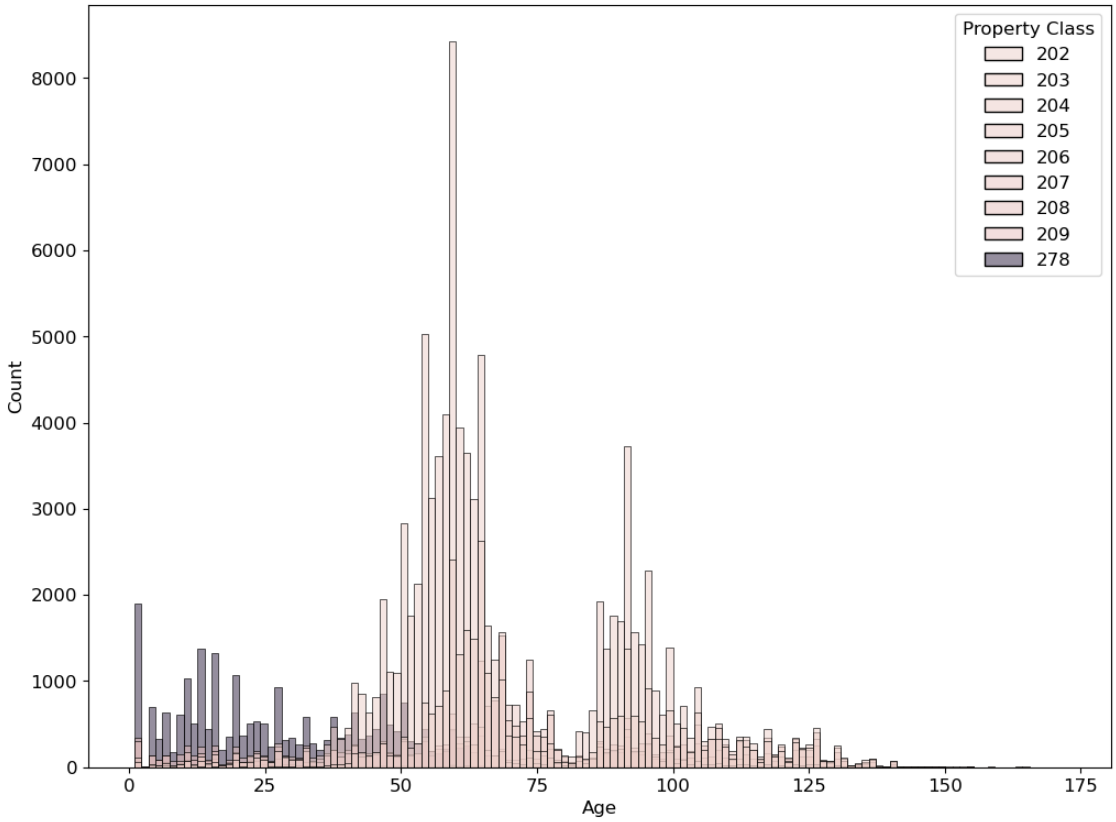
```
In [12]: sns.kdeplot(data=initial_data, x='Age')
```

```
Out[12]: <Axes: xlabel='Age', ylabel='Density'>
```




```
In [13]: sns.histplot(initial_data, x='Age', hue='Property Class')
```

```
Out[13]: <Axes: xlabel='Age', ylabel='Count'>
```


Property Class
202
203
204
205
206
207
208
209
278
Count
Age
0
1000
2000
3000
4000
5000
6000
7000
8000
0
25
50
75
100
125
150
175

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.

How does the race of the a property owner correlate with their average property sale value?

- I would create a bar graph plotting the mean sale-price of properties on the y-axis against the different races recorded in the dataset on the x-axis. For every race/ethnicity given in the dataset, I would calculate a mean property value for the properties owned by people of that race.

0.5 Question 2a

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

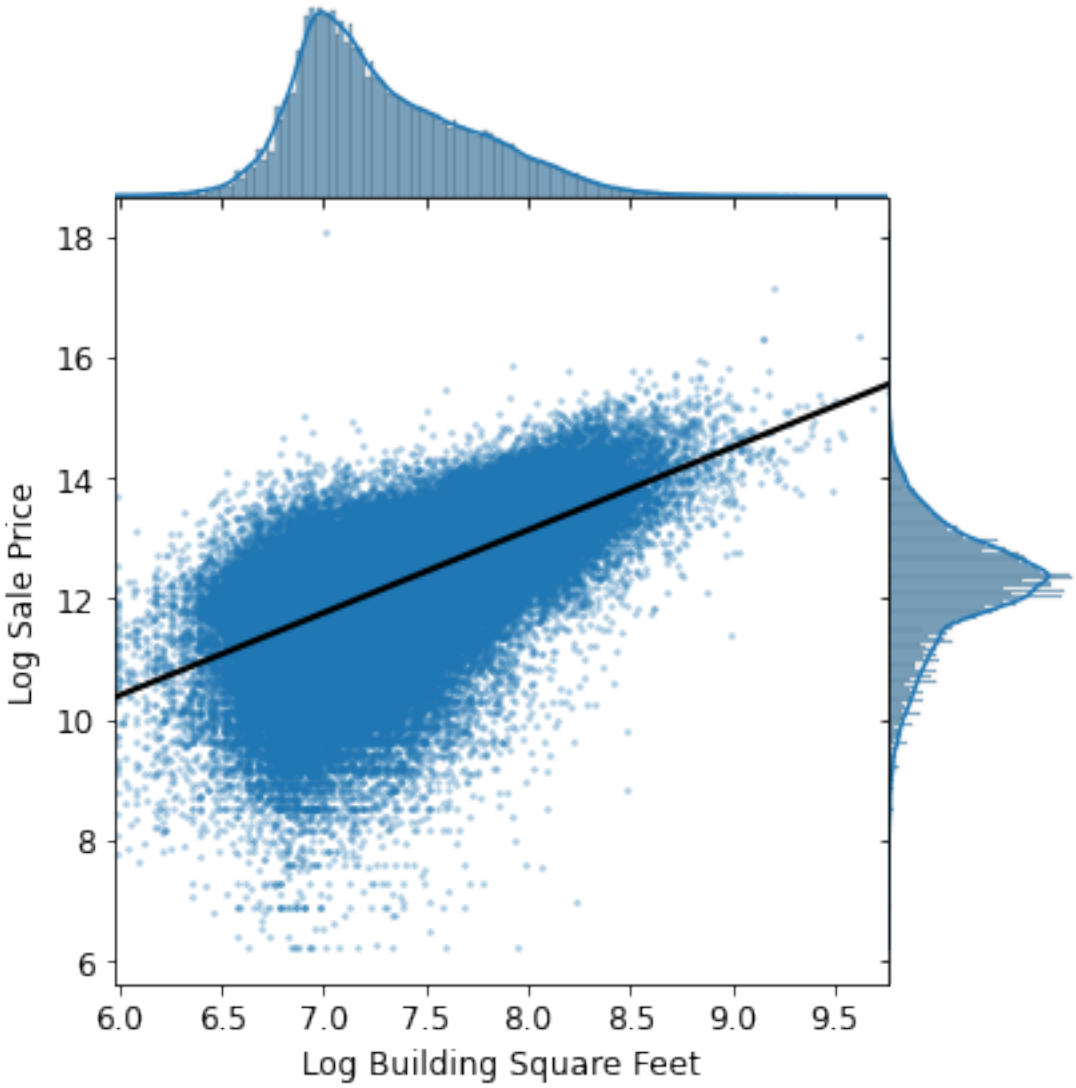
Because the maximum sale price is \$70 million (!), the plotting distribution becomes incredibly right-skewed. The graphs must accomodate for the one outlier of \$70 million, and thus the graph becomes extremely difficult to interpret. In order to overcome it, one could log transform the Sale Prices. This would bring the outlier of \$70 million closer to the rest of the data points, while not significantly disturbing the smaller data values (relative to the \$70 million). The new plots are shown below.

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 does not look like the perfect feature to include in our linear model because it doesn't appear to have a linear relationship with Log Sale Price. One can see the curve is bulging to the upper left, forming a kind of arc like in the upper left quadrant of a circle. The relationship does not appear to be linear, or at least is not linear enough for my liking. Therefore, it does not make a good candidate for a feature for our linear model.

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`.

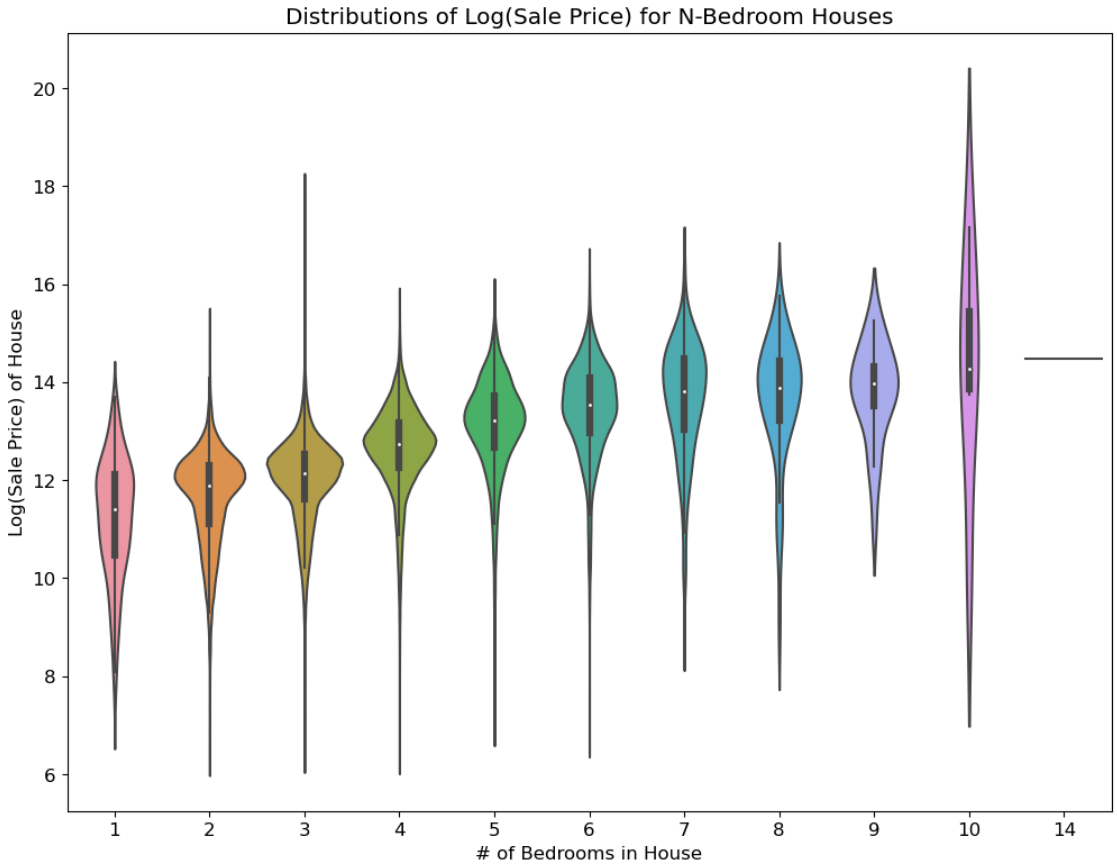
Hint: A direct scatter plot of the `Sale Price` against the number of rooms for all of the households in our training data might risk overplotting.

In [37]:
```
# fig, ax = plt.subplots()
# sns.regplot(data=training_data, x='Bedrooms', y='Log Sale Price', ax=ax)
# sns.lmplot(data=training_data, x='Bedrooms', y='Log Sale Price', ax=ax)
```

In [38]:
```
# sns.boxplot(data=training_data, x='Bedrooms', y='Log Sale Price')
```

In [39]:
```
fig, ax = plt.subplots()
sns.violinplot(data=training_data, x='Bedrooms', y='Log Sale Price', ax=ax)
ax.set_title('Distributions of Log(Sale Price) for N-Bedroom Houses')
ax.set_ylabel('Log(Sale Price) of House')
ax.set_xlabel('# of Bedrooms in House')
```

Out[39]:
```
Text(0.5, 0, '# of Bedrooms in House')
```



In [40]: *# sns.boxenplot(data=training_data, x='Bedrooms', y='Log Sale Price')*