
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 full description of a house sold in Cook County. It contains all the details of the house and its neighborhood.

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.

I think housing department, internal revenue service and police department would like to know the data. Housing department and internal revenue service need the data for understanding the housing market and the tax. Police department would need the data to evaluate the neighborhood.

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.

'Site Desirability' seems to contain demographic information. I think it's an indicator of how popular the house is, e.g. how long it stays on the market or how many offers the house receive, etc.

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: I would like to understand which parameter is mostly impacting the final sale price in a given year. For such a purpose we would need to draw sns.replot(s) for all the parameters vs the 'Sale Price'.

Question 2: I would like to understand how the 'Sale Price' grows with 'Sale Year' in different 'Neighborhood Code'. It would also be interesting to see whether the neighborhood population increased or decreased over the years.

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 the most expensive house is way above the scale of the median house sale price. Hence by using the max sale price as the upper limit for the scale bar, most of the data is squeezed to the left corner.

One can either use 75% percentile as the x_max, or log the x axis to uniformly distribute the datapoints.

```
In [101]: training_data['Sale Price'].describe()
```

```
Out[101]: count    2.047920e+05
          mean     2.451646e+05
          std      3.628694e+05
          min      1.000000e+00
          25%      4.520000e+04
          50%      1.750000e+05
          75%      3.120000e+05
          max      7.100000e+07
          Name: Sale Price, dtype: float64
```

0.6 Question 2b

To zoom in on the visualization of most households, we will focus only on a subset of `Sale Price` for this assignment. In addition, it may be a good idea to apply log transformation to `Sale Price`. In the cell below, reassign `training_data` to a new dataframe that is the same as the original one **except with the following changes**:

- `training_data` should contain only households whose price is at least $500.
- `training_data` should contain a new `Log Sale Price` column that contains the log-transformed sale prices.

Note: This also implies from now on, our target variable in the model will be the log-transformed sale prices from the column `Log Sale Price`.

Note: You should **NOT** remove the original column `Sale Price` as it will be helpful for later questions.

To ensure that any error from this part does not propagate to later questions, there will be no hidden test here.

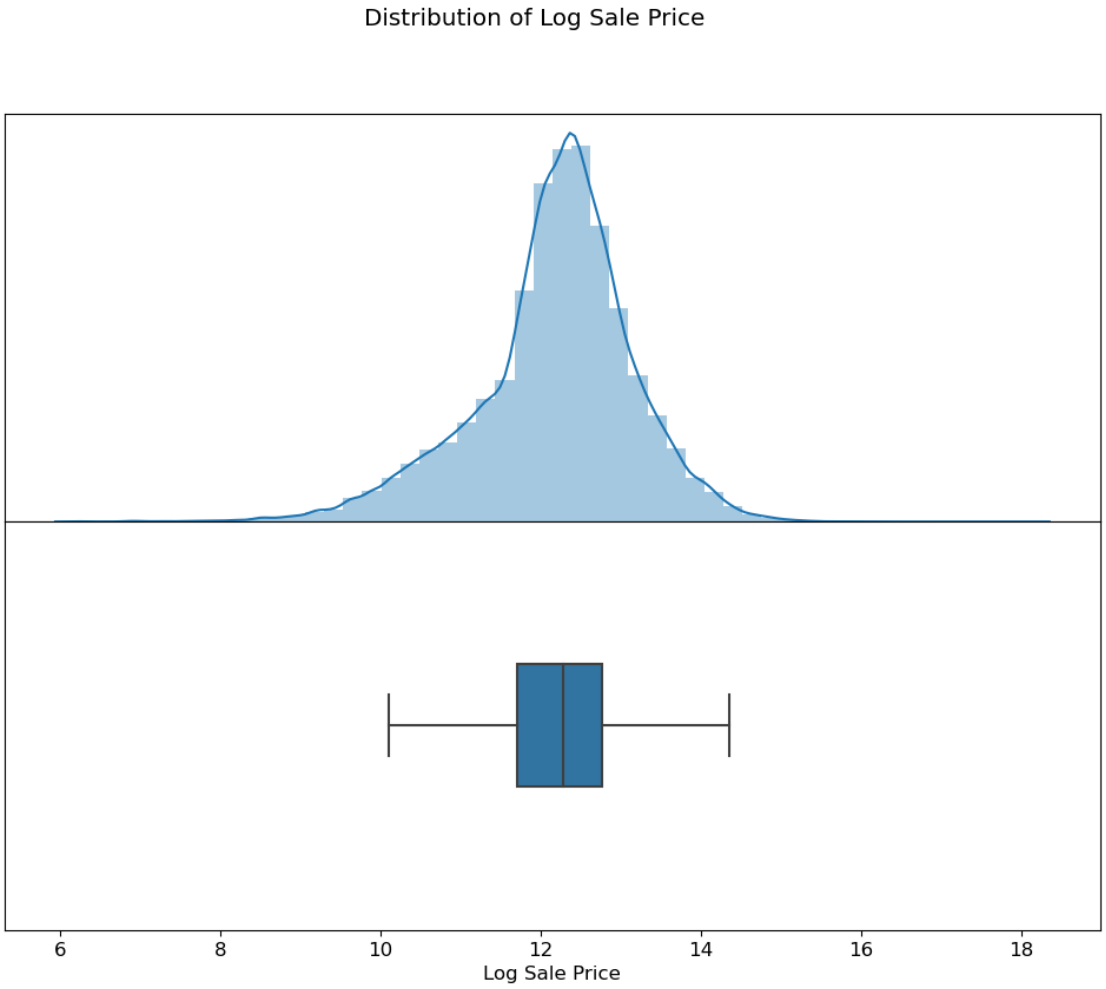
```
In [102]: training_data = training_data[training_data['Sale Price']>=500]
          training_data['Log Sale Price'] = np.log(training_data['Sale Price'])
```

```
In [103]: grader.check("q2b")
```

```
Out[103]: q2b results: All test cases passed!
```

Let's create a new distribution plot on the log-transformed sale price.

```
In [104]: plot_distribution(training_data, label='Log Sale Price');
```



0.7 Question 3a

Is the following statement correct? Assign your answer to `q3statement`.

```
"At least 25% of the houses in the training set sold for more than $200,000.00."
```

The provided test for this question do not confirm that you have answered correctly; only that you have assigned each variable to `True` or `False`.

```
In [105]: # This should be True or False
          q3statement = True
```

```
In [106]: grader.check("q3a")
```

```
Out[106]: q3a results: All test cases passed!
```

0.8 Question 3b

Next, we want to explore if any there is any correlation between `Log Sale Price` and the total area occupied by the household. The `codebook.txt` file tells us the column `Building Square Feet` should do the trick – it measures "(from the exterior) the total area, in square feet, occupied by the building".

Before creating this `jointplot` however, let's also apply a log-transformation to the `Building Square Feet` column.

In the following cell, create a new column `Log Building Square Feet` in our `training_data` that contains the log-transformed area occupied by each household.

You should NOT remove the original `Building Square Feet` column this time as it will be used for later questions.

To ensure that any errors from this part do not propagate to later questions, there will be no hidden tests here.

```
In [107]: training_data['Log Building Square Feet'] = np.log(training_data['Building Square Feet'])
```

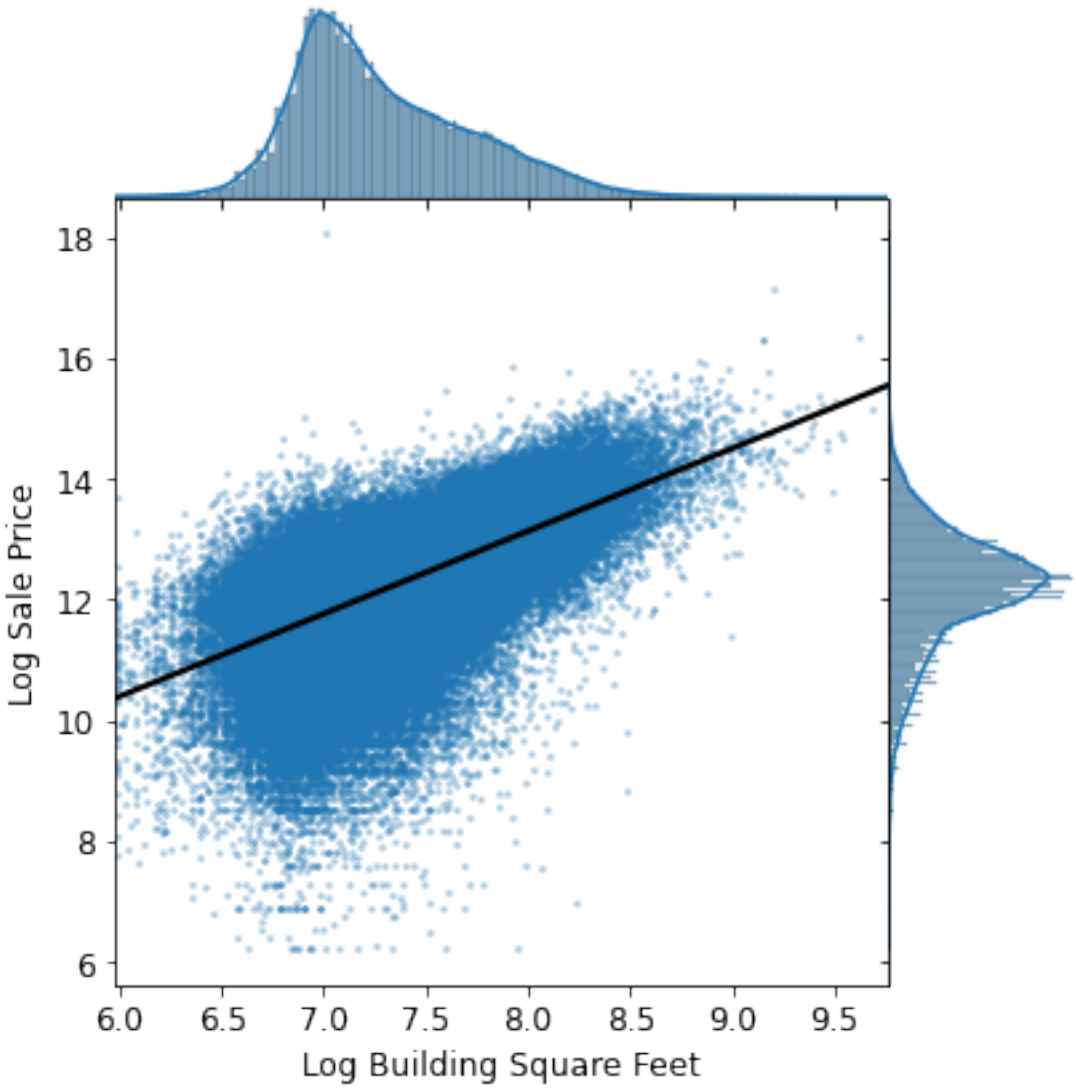
```
In [108]: grader.check("q3b")
```

```
Out[108]: q3b results: All test cases passed!
```

0.9 Question 3c

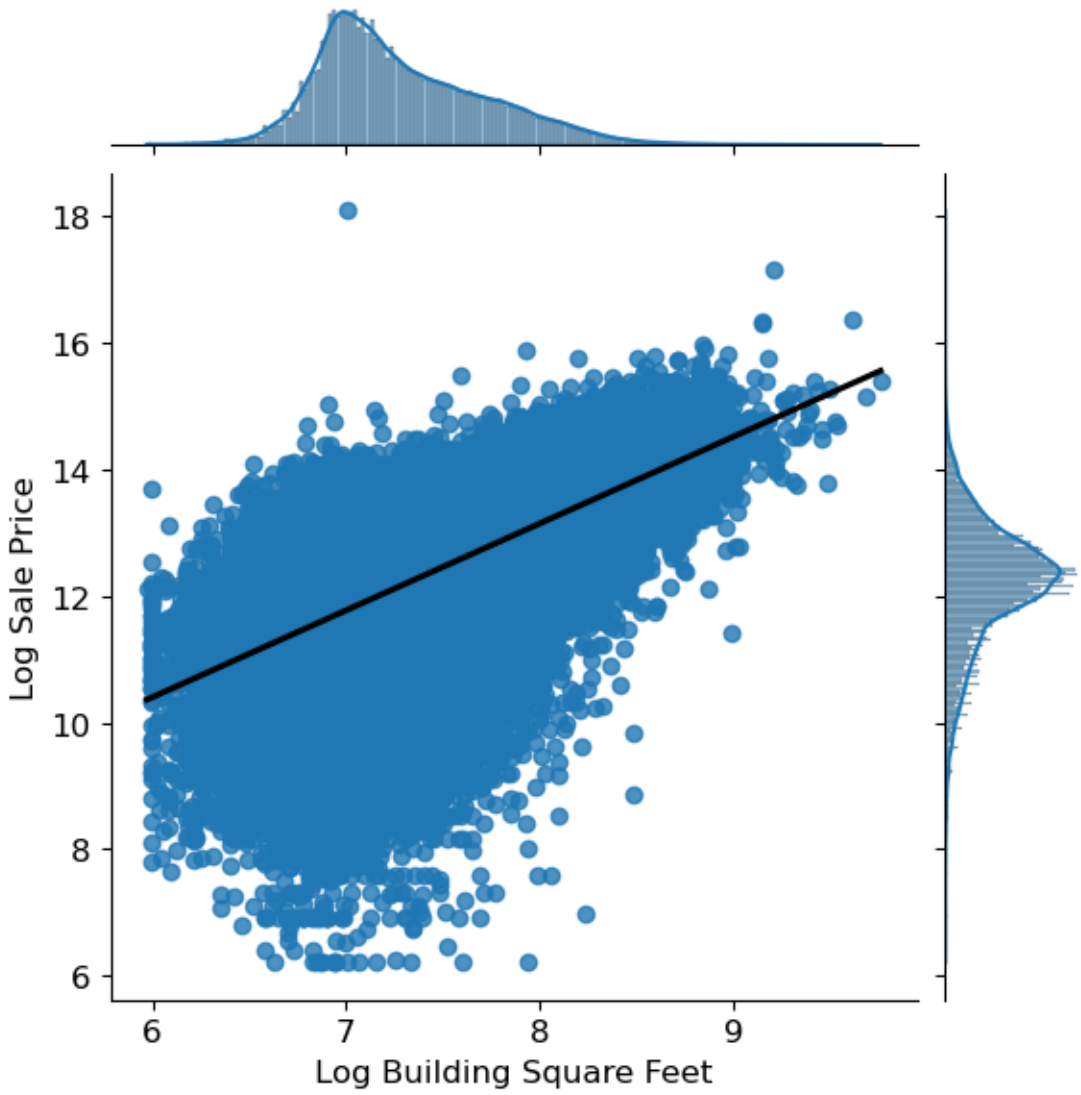
As shown below, we created a `jointplot` with `Log Building Square Feet` on the x-axis, and `Log Sale Price` on the y-axis. In addition, we fit a simple linear regression line through the bivariate scatter plot in the middle.

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?




```
In [109]: sns.jointplot(data = training_data, x = 'Log Building Square Feet', y = 'Log Sale Price', kin
                       joint_kws={'line_kws':{'color':'black'}})
```

```
Out[109]: <seaborn.axisgrid.JointGrid at 0x7f34fb1379a0>
```



I think `log building square feet` can serve as a good candidate feature for our model. It's largely correlated with the `log sale price` and shows a good agreement in trend.

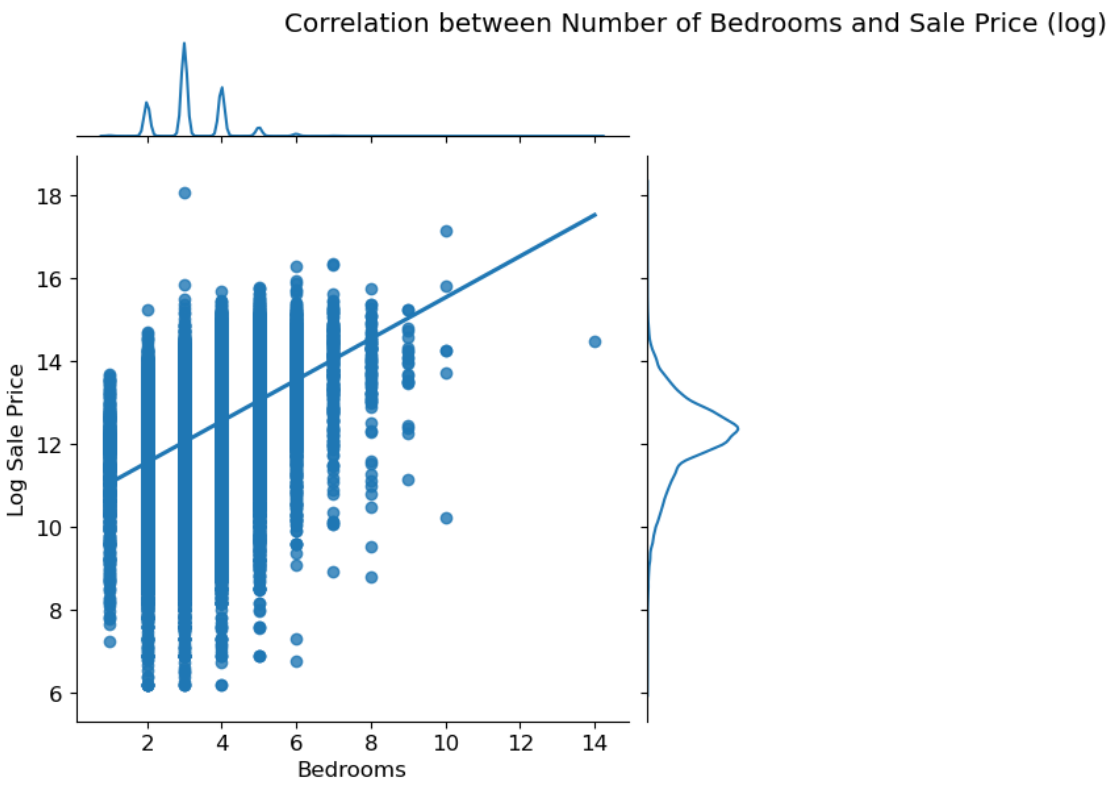
0.10 Question 5c

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]: g=sns.JointGrid(data = training_data, x='Bedrooms', y='Log Sale Price')
          g.plot(sns.regplot, sns.kdeplot)
          plt.title('Correlation between Number of Bedrooms and Sale Price (log)', y = 1.2)
```

```
Out[117]: Text(0.5, 1.2, 'Correlation between Number of Bedrooms and Sale Price (log)')
```



0.11 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?

Although the plot looks cleaner, there doesn't seem to be a correlation between `log sale price` and the `neighborhood code`. Different neighborhood codes share almost the same median log sale prices. I think instead of neighborhood codes or sale amount in the neighborhood, there might be other factors.