
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?

The granularity of the dataset is at the individual property sale level. Each row represents a single housing transaction in Cook County, Illinois, containing detailed information about the property, such as location and sale price.

0.2 Question 1b

Why was this data collected? For what purposes? By whom?

You should watch Lecture 13 before attempting this question.

This data was collected by the Cook County Assessor's Office to estimate the market value of properties across Cook County, Illinois. Its purpose is to determine fair property taxes to allow homeowners to contribute appropriately to public services. Each property's characteristics, like its size, location, and condition, help predict its value. Also, the data allows the office to monitor accuracy and address inequities in the assessment process.

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.

My first question is "How does a Property's Land Square Feet relate to its Sale Price? TO answer this question, I would create a scatter plot of Land Square Feet on the x-axis versus the Sale Price on the y-axis to see whether larger lots tend to sell for higher prices. My second question is"How does the Age of the property affect the sale price across different property classes? To answer this, I would make separate scatter plots of each property class that would visualize Age with Sale Price to see if older buildings generally sell for less within each category.

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.

A new question I would ask is "How does the homeowner's annual income relate to the Sale Price of their property across different Neighborhood Codes?" To answer this, I would create a scatter plot of annual income on the x-axis and Sale Price on the y-axis, with the Neighborhood Code as the hue, to explore whether income levels correlate with property values and reveal potential fairness disparities across neighborhoods.

0.5 Question 1e

Look at `codebook.txt` to see some of the unique regional features CCAO utilizes, such as `O'Hare Noise`. Now imagine you were in charge of predicting the `Sale Price` of houses in **your hometown** (your actual real life hometown/city - not the data provided). Propose a feature that you would want to collect specific to your location and hypothesize why it might be useful in predicting the sale price of houses.

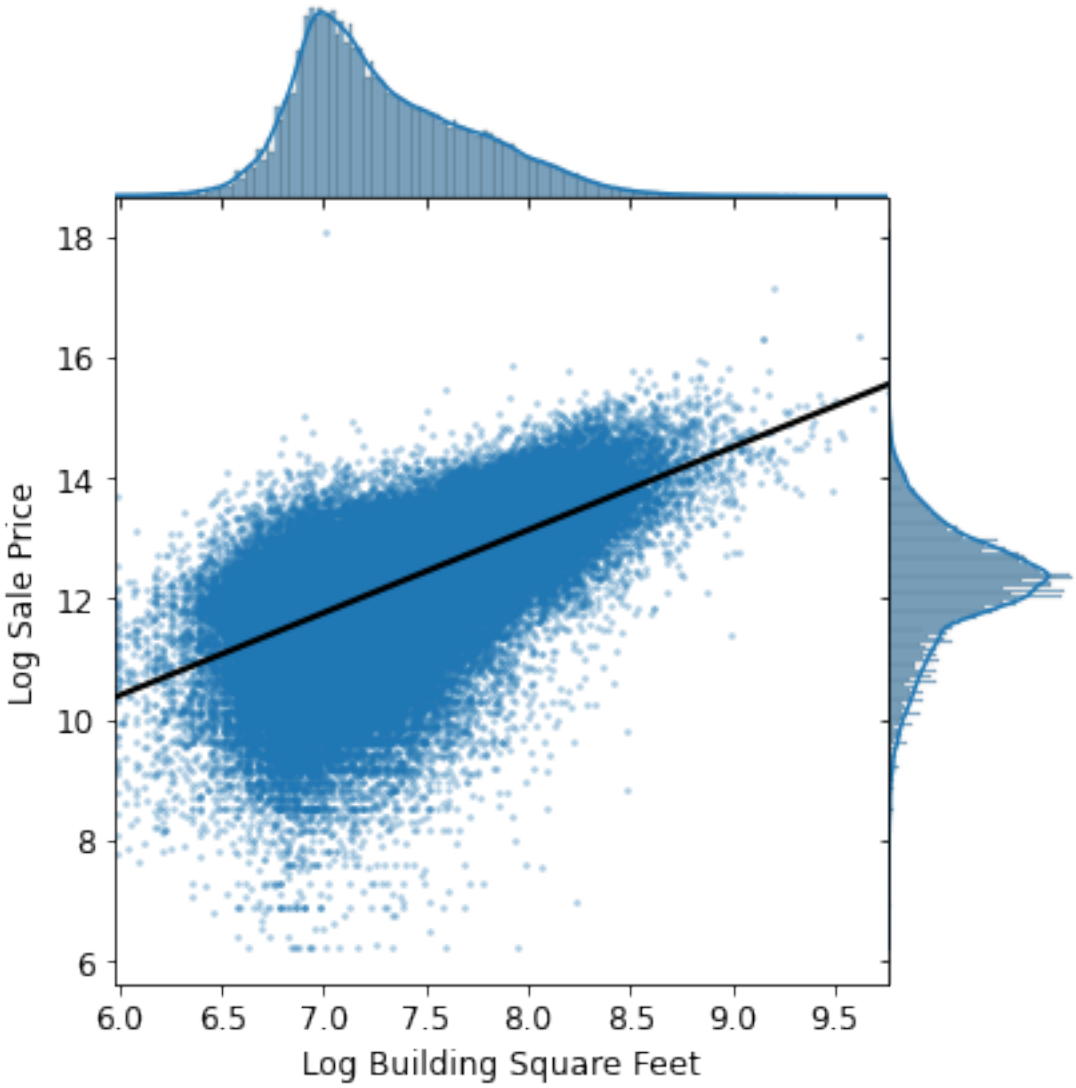
If I were predicting house prices in my hometown of Los Angeles, I would want to collect a feature that measures the distance to major entertainment and cultural hotspots, such as downtown LA, Hollywood, or Santa Monica. Proximity to these places often raises housing demand due to job availability, nightlife, and other benefits. This feature could help explain the variation in sale prices across different Los Angeles neighborhoods.

0.6 Question 3b

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, Log Building Square Feet would make a good candidate feature for predicting Log Sale Price. The scatter plot shows a clear positive linear relationship. As the building size increases, the sale price tends to increase as well. The points are pretty concentrated around the regression line, which suggests a consistent trend rather than random variation. This indicates that larger buildings generally sell for higher prices, making this variable meaningful and predictive for our model.

0.7 Question 5c

Create a visualization that clearly and succinctly shows if there exists an association between `Bathrooms` 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 `Bathrooms`.

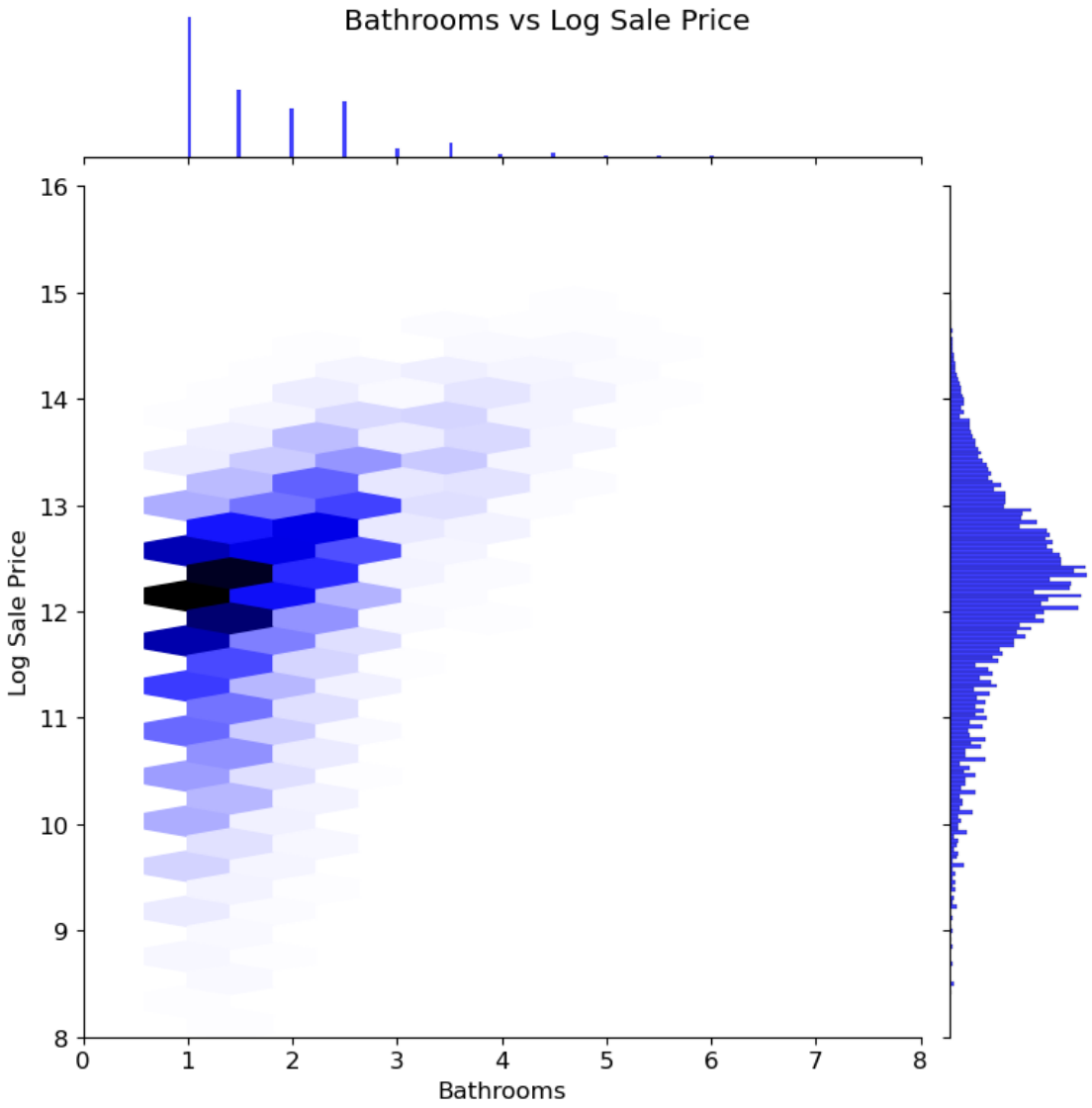
Hint 1: A direct scatter plot of the `Sale Price` against the number of rooms for all of the households in our training data might risk overplotting. Try it out and see why causes overplotting.

Hint 2: Take a closer look at the variable `Bathrooms`. What variable type is `Bathrooms` and what are the possible values `Bathrooms` can take? What visualizations might be good at showing this?

```
In [33]: g = sns.jointplot(
             data=training_data,
             x='Bathrooms',
             y='Log Sale Price',
             kind='hex',
             color='blue',
             height=8
         )

         g.ax_joint.set_xlim(0, 8)
         g.ax_joint.set_ylim(8, 16)

         plt.suptitle("Bathrooms vs Log Sale Price")
         plt.show()
```


Bathrooms vs Log Sale Price
Log Sale Price
Bathrooms
16
15
14
13
12
11
10
9
8
0
1
2
3
4
5
6
7
8

0.8 Question 8: Open-Ended Question

Welcome to another **open-ended question**.

> If you have any feedback on this open-ended question, or any other homework question in Data 100, we encourage you to share your thoughts using the content feedback form. You can also post to Ed.

Grading on open-ended questions is simple: **Clear evidence of thoughtfulness and effort will always receive full credit**. If your response is especially well-developed or creative, we may ask for permission to share it with the rest of the class so others can be inspired by your work! Underdeveloped ideas will receive half credit. Trivial or missing responses will receive no credit. We expect the vast majority of students to receive full credit.

SETUP: As mentioned in previous lectures, RMSE is a very powerful error metric. However, in lecture 13 we learned that RMSE is not the best choice when evaluating the *fairness* of a property appraisal system and can be quite limiting. You can read about this in more detail here or rewatch lecture 13. In short, RMSE fails to tell you:

1. The distribution of the errors
2. The sign (+/-) of the errors
3. The relative size of the errors

These aspects can be crucial when we are performing data analysis and developing a predictive model ourselves in this project.

TASK: Your task is to design your own custom error metric that may outperform RMSE in some of the aspects above to help you build a better machine learning model. Later, in Project A2 you will apply this custom error metric and compare it against RMSE

Your answer should consist of the following: 1. **Define a mathematical formula** for your custom error metric that penalizes unfair predictions 2. **Explain your reasoning** for why you think this metric captures fairness better than standard metrics like MSE or MAE 3. **Identify a potential counterargument** against your custom error metric. Why it might not be ideal in this context?

In the following cell (feel free to use LaTeX in combination with Markdown syntax), define a custom error metric that you believe would better capture fairness concerns in housing price prediction.

Important exception to existing course policies: **FOR THIS QUESTION ONLY** you are allowed to use large-language models (LLMs), like Gemini or ChatGPT. **However**, we strongly recommend thinking through this question on your own first and generating some different options. Only then you should use an LLM to discuss your possible metrics further. **If you copy-and-paste default output from an LLM on this question, there is a good chance that your submission will look identical or near-identical to many other students**. While we expect many answers to this question to have similarities, obvious default output will receive no credit. Spend time thinking about the presentation of your results.

Hint: If you are completely lost, you might consider metrics that weight errors differently based on property characteristics, or that penalize systematic biases across different groups.

> **Disclaimer**: As Data Science students, you should be aware of important limitations and broader considerations when it comes to the use of LLMs. - LLMs do not guarantee factual accuracy and they are known to hallucinate (generate fabricated or misleading information). - LLMs are trained on large datasets that can reflect and reproduce biases in race, gender, culture, and ideology. - The use of LLMs may involve the sharing of sensitive and personal information.

> **IMPORTANT**: If you have any questions, please read through the **FAQS** first. If you can't find the answer to your question there, feel free to ask your question on Ed.

A custom error metric I have created, which penalizes unfair predictions, will combine both the Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) to balance absolute and relative prediction accuracy.

$$FairnessError = \alpha \times MAPE + (1 - \alpha) \times MAPE$$

where

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

and

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{\hat{y} - y_i}{y_i}$$

and

$$\alpha \in [0, 1]$$

which represents the trade-off between the two. This metric would capture fairness better than RMSE because it considers the absolute errors of the MAE and the proportional errors from the MAPE to prevent high-value homes from affecting the loss function. By scaling errors relative to each home's value, it penalizes disproportionate mistakes on lower-priced properties, promoting socioeconomic fairness. One thing to note, however, is that the MAPE can overreact to very small $ y_i$ values, and the chosen α may introduce subjectivity.