
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 in the dataset corresponds to a distinct property transaction. The dataset includes various details about the property, such as its physical characteristics (e.g., size, structure, construction materials), location (e.g., neighborhood, township, geographic coordinates), and transaction-related information (e.g., sale price, sale date, deed number).

0.2 Question 1b

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

You should watch Lecture 15 before attempting this question.

The data was collected by the Cook County Assessor's Office (CCAO) to assess property values for taxation purposes. The goal was to ensure that property taxes were assigned fairly and equitably across different neighborhoods. This data helps in building valuation models to estimate property values, even for homes that have not been recently sold, improving transparency and reducing bias in tax assessments.

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.

1. How does the sale price of properties vary across different neighborhoods in Cook County? I would create a box plot of sale price (SALE PRICE) by neighborhood (NEIGHBORHOOD CODE) to visualize price distributions and identify disparities in property values.
2. Is there a relationship between land square footage and sale price? I would create a scatter plot of land square feet (LAND SQUARE FEET) against sale price (SALE PRICE) to observe whether larger properties tend to have higher prices. I would also calculate the correlation coefficient between these two variables to quantify the strength of the relationship

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.

Is there a relationship between a homeowner's annual income and the sale price of their property in Cook County?

Analytical approach: - I would create a scatter plot of annual income (from the demographic dataset) against sale price (SALE PRICE) to observe whether higher-income individuals tend to own more expensive properties. - I would calculate the correlation coefficient to measure the strength and direction of the relationship. - Additionally, I could perform a regression analysis, with sale price as the dependent variable and annual income as an independent variable, to determine the effect of income on property values while controlling for other factors.

This analysis would help assess economic disparities in homeownership and pricing trends in Cook County.

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.

Feature Proposal: Proximity to Public Transportation (MRT/LRT/TransJakarta Stations in Jakarta)

Hypothesis: In Jakarta, access to reliable public transportation significantly impacts property values. Houses located near MRT, LRT, or TransJakarta stations are often more desirable due to reduced commuting time, convenience, and lower transportation costs. Adding a feature that measures the distance from a property to the nearest transit hub could improve sale price predictions.

Justification: - Properties closer to major transit hubs tend to have higher demand, leading to higher sale prices. - Government infrastructure projects, such as TOD (Transit-Oriented Development) zones, often lead to property appreciation. - This feature would be particularly useful in urban areas where public transport accessibility is a key selling point.

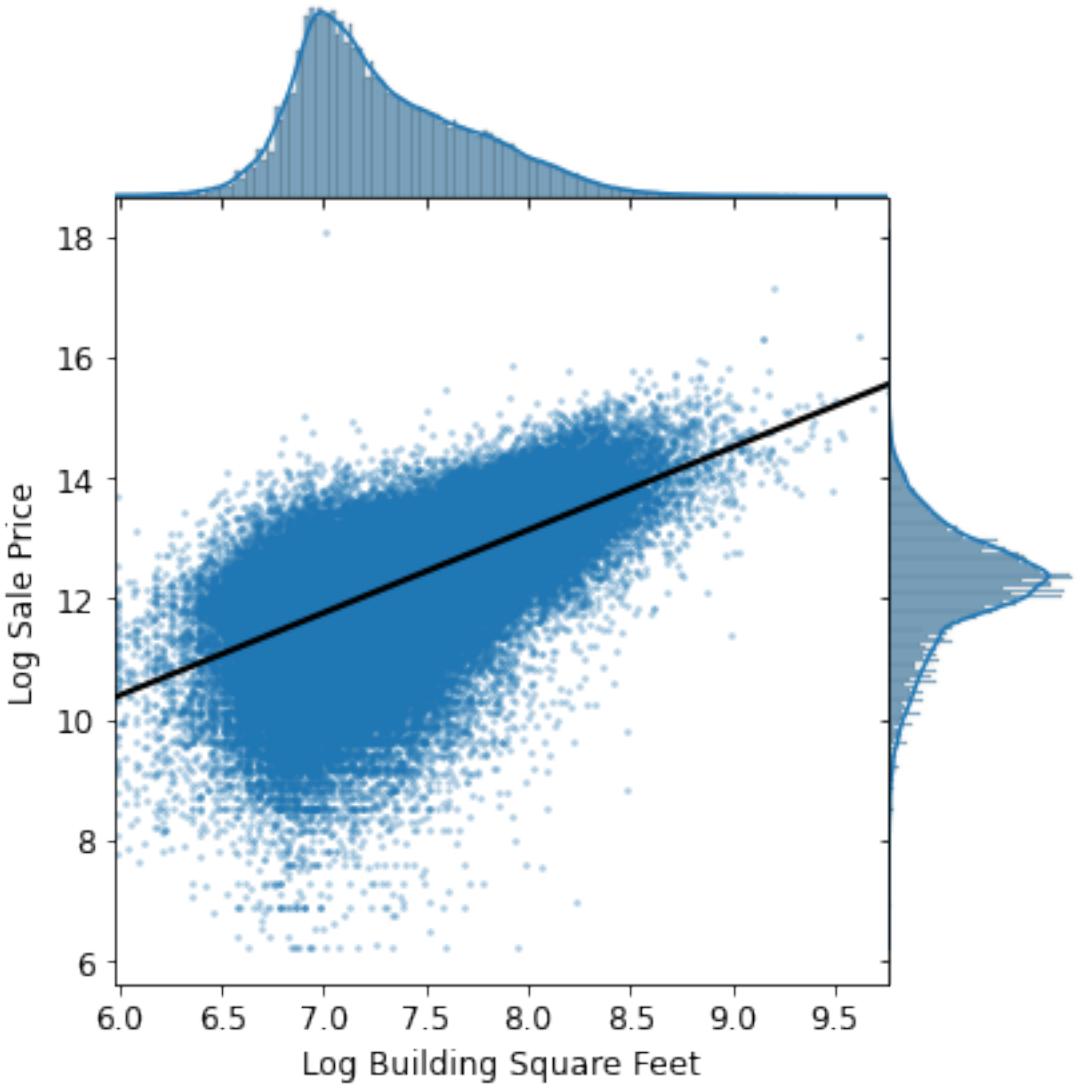
This information could be collected using GIS data or Google Maps API to calculate the walking or driving distance from each property to the nearest station.

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, the jointplot suggests that Log Building Square Feet is a reasonable candidate for inclusion in our model. There is a noticeable positive correlation between Log Building Square Feet and Log Sale Price, meaning that larger properties tend to have higher sale prices. While the data exhibits some variance, the overall trend follows a linear relationship, as indicated by the fitted regression line.

For a feature to be valuable in a predictive model, it should have a consistent and interpretable relationship with the target variable. Here, Log Building Square Feet aligns with economic intuition—larger buildings typically command higher prices due to increased utility, desirability, and cost of materials. Given this, it makes sense to include it in the model to improve its predictive power. However, further exploration—such as checking for outliers, multicollinearity with other variables, and nonlinearity—would be necessary before finalizing its inclusion.

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: 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 [86]: import matplotlib.pyplot as plt
         import seaborn as sns

         plt.figure(figsize=(14,6))

         filtered_data = training_data[training_data['Bathrooms'] <= 10]

         sns.boxplot(x='Bathrooms', y='Log Sale Price', data=filtered_data)

         plt.xlabel('Bathrooms')
         plt.ylabel('Log Sale Price')
         plt.title('Association between Bathrooms and Log Sale Price')

         plt.xticks(ticks=sorted(filtered_data['Bathrooms'].unique()), rotation=0)

         plt.show()
```