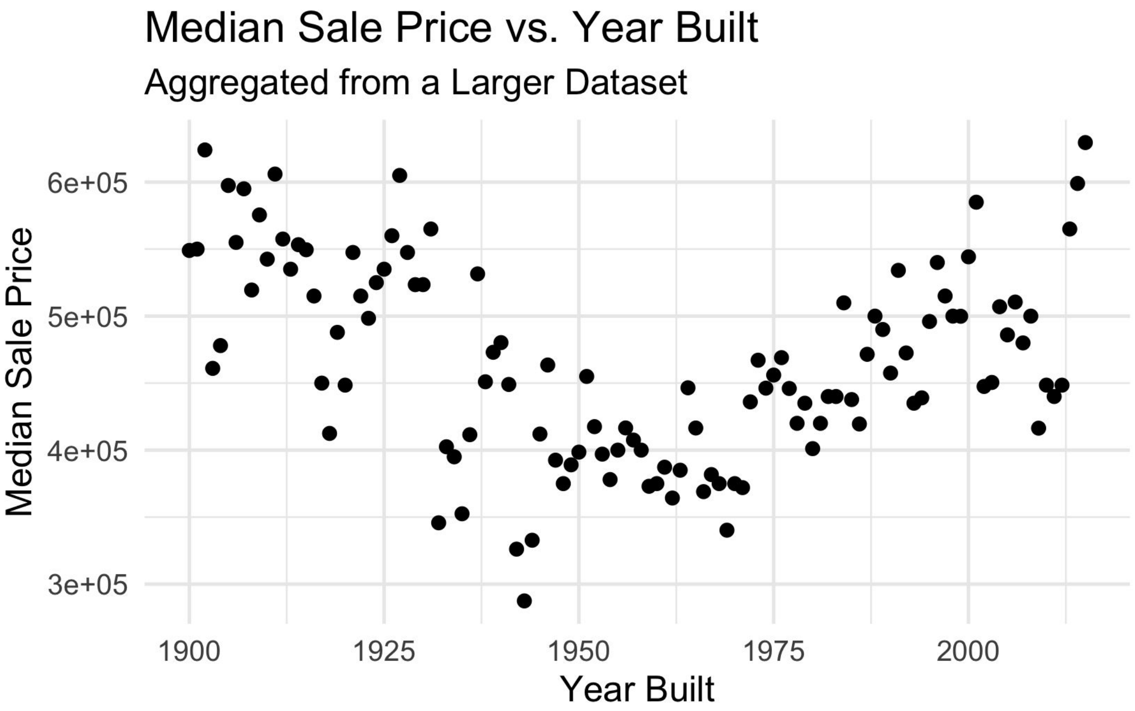
MACHINE LEARNING SERIES, #1

Intro to Statistical Learning

AN INTRODUCTION TO STATISTICAL MODELS AND THE MODELING PROCESS

DS Collab, Winter 2025
Presented by: Ethan P. Marzban

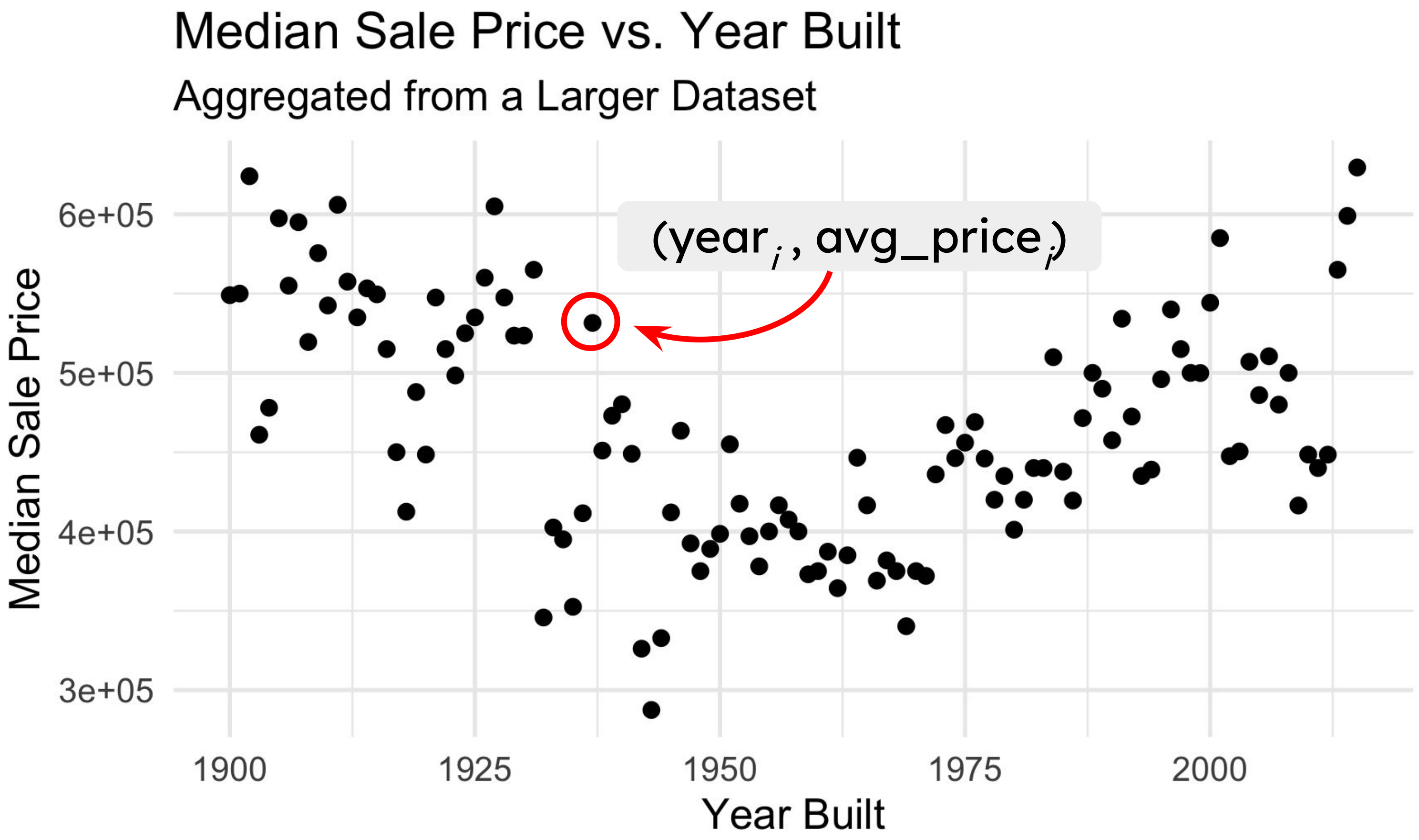
Example



- Median sale price of homes in King County, WA; averaged by year
- Original dataset accessed from Kaggle; aggregation performed using tidyverse commands
- **Check our Intuition:** does the “dip” in the plot make sense?

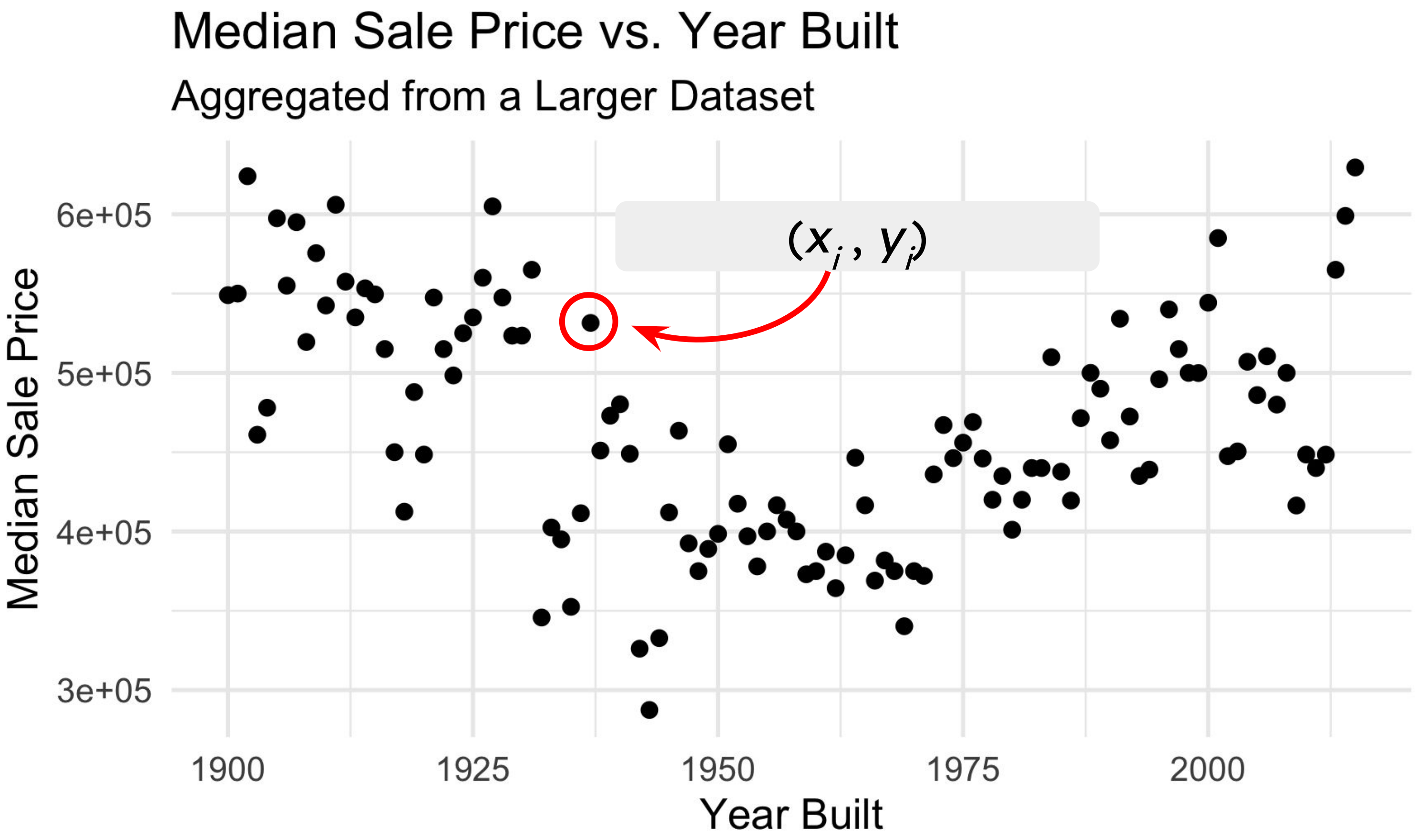
- This is a **bivariate** plot: it contains information on two variables.
  - **Check your Understanding:** what are the types (continuous, discrete, ordinal, nominal) of these two variables?

Example



Median Sale Price vs. Year Built
Aggregated from a Larger Dataset
(year$_i$, avg_price$_i$)
Median Sale Price
6e+05
5e+05
4e+05
3e+05
1900
1925
1950
1975
2000
Year Built

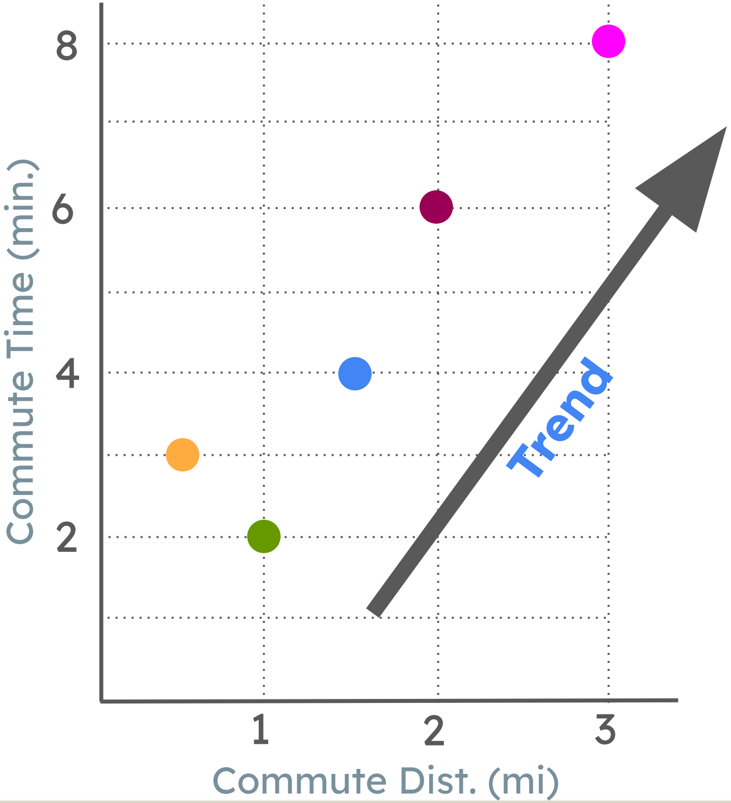
Example


Median Sale Price vs. Year Built
Aggregated from a Larger Dataset
Median Sale Price
6e+05
5e+05
4e+05
3e+05
1900
1925
1950
1975
2000
Year Built
(x_i, y_i)

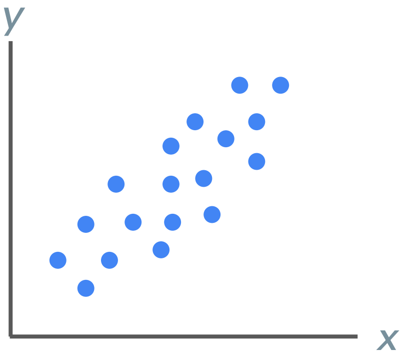
Review: Scatterplots

- **Scatterplots** are the ideal visualization for **bivariate numerical data** (i.e. data whose observations are pairs of numbers)
  - Each observation gets mapped to a point on the cartesian plane

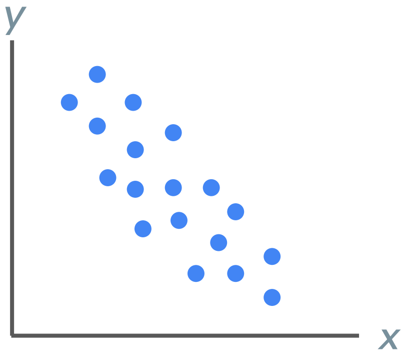
Commute Dist.	Commute Time
0.5	3
1	2
1.5	4
2	6
2.5	8



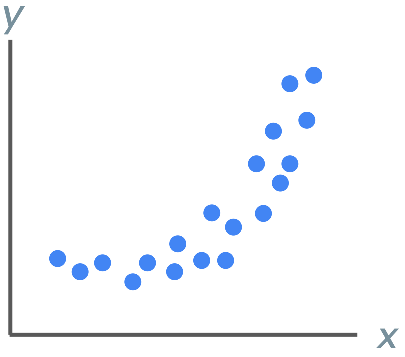
Review: Trend



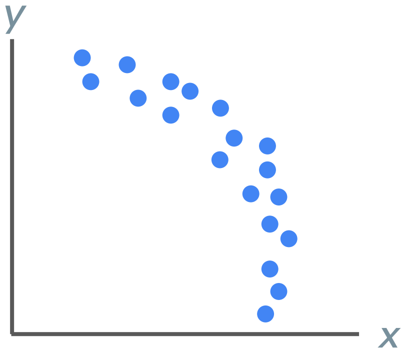
- **Positive**: as *x* increases, *y* also increases
- **Linear**: the rate of change is roughly constant



- **Negative**: as *x* increases, *y* decreases
- **Linear**: the rate of change is roughly constant

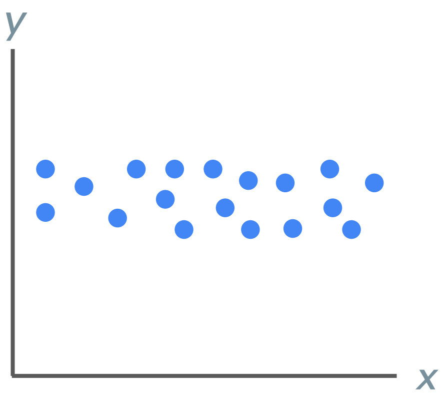


- **Positive**: as *x* increases, *y* also increases
- **Nonlinear**: the rate of change is nonconstant



- **Negative**: as *x* increases, *y* decreases
- **Nonlinear**: the rate of change is nonconstant

Review: Trend



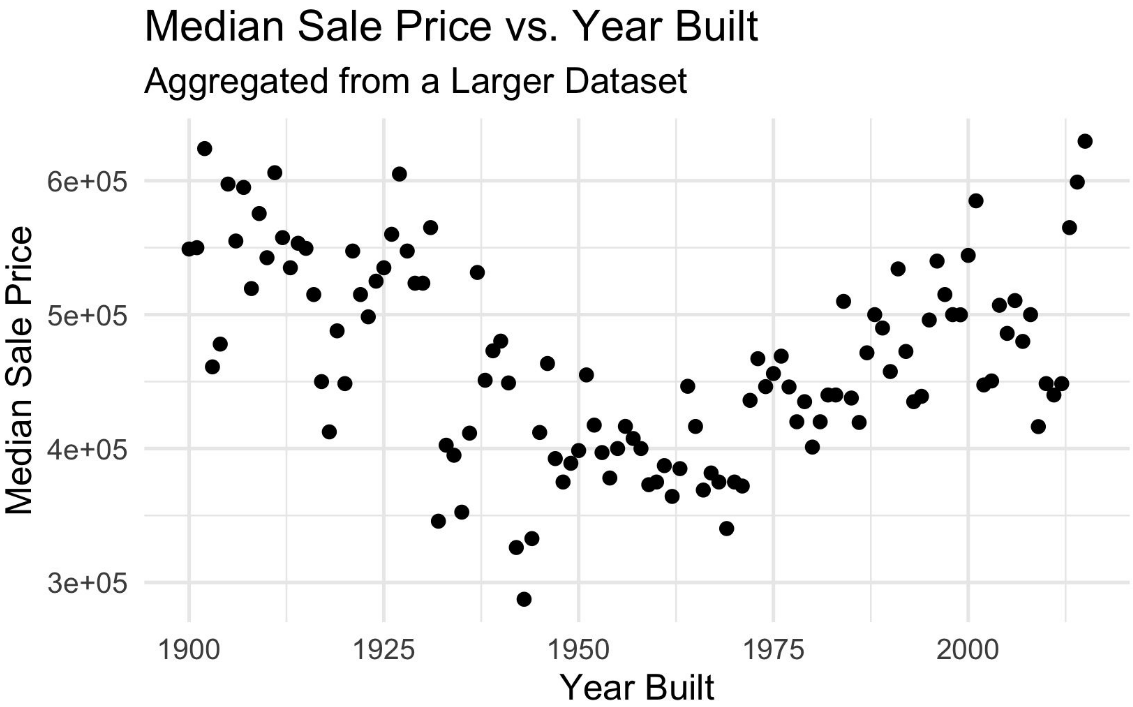
- No trend.
  - As *x* increases, *y* appears to stay relatively constant.

- Another way to describe the findings of a scatterplot is in terms of the **association** between the variables being compared.
  - For instance, if the scatterplot of *y* vs. *x* displays a positive linear trend, we would say that *x* and *y* have a positive linear association, or that *x* and *y* are positively linearly associated.

- If the scatterplot displays no trend, then we say that the variables being compared do not have an association.

- Measures like **correlation** seek to quantify the strength of a linear relationship between two variables, but do not really provide information on the *structure* of the relationship

Example



- **Test yourself:** classify the type of trend that is apparent in our Average Sales Price vs. Year Built scatterplot

Introduction to Modeling

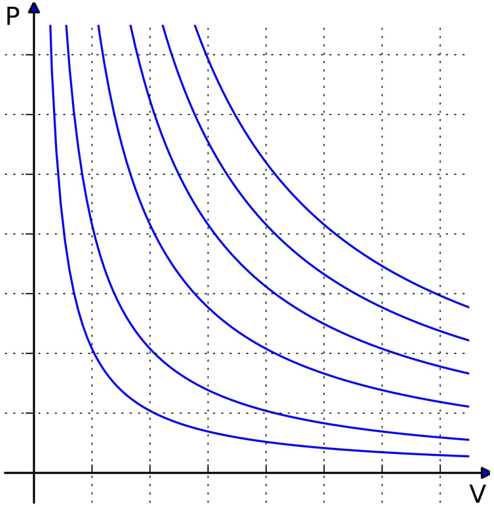
Models

- So, what’s a model?
- We can think of a **model** as a mathematical or *idealized* representation of a system.
  - E.g. a weather model seeks to represent what we think the weather tomorrow will be like
- Two types: **deterministic** and **probabilistic**.
- Deterministic models govern the way the world works.

Ideal Gas Law: Seeks to model the behavior of gas under certain conditions.

States that the product of the pressure and volume of the gas is proportional to the product of the amount of gas and the temperature.

$$pV = nRT$$



Source: https://en.wikipedia.org/wiki/Gas_constant

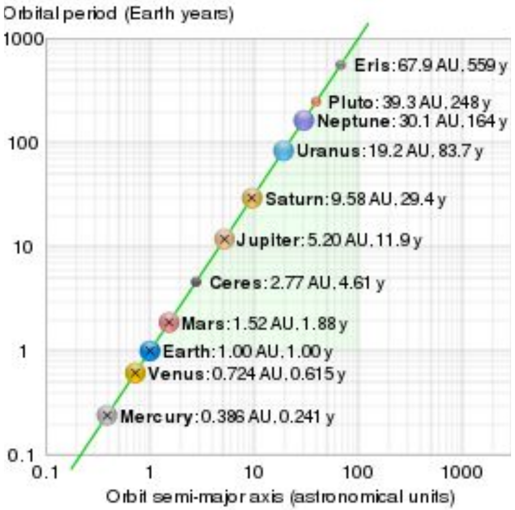
Isotherms of an ideal gas for different temperatures. The curved lines are rectangular hyperbolae of the form y = a/x. They represent the relationship between pressure (on the vertical axis) and volume (on the horizontal axis) for an ideal gas at different temperatures: lines that are farther away from the origin (that is, lines that are nearer to the top right-hand corner of the diagram) correspond to higher temperatures.

Models

- So, what’s a model?
- We can think of a **model** as a mathematical or *idealized* representation of a system.
  - E.g. a weather model seeks to represent what we think the weather tomorrow will be like
- Two types: **deterministic** and **probabilistic**.
- Deterministic models govern the way the world works.

Kepler’s Third Law of Planetary Motion: Seeks to model the behavior of planets.

The ratio of the square of an object’s orbital period with the cube of the semi-major axis of its orbit is the same for all objects orbiting the same primary.



Source: https://en.wikipedia.org/wiki/Kepler%27s_laws_of_planetary_motion#Third_law

Log-log plot of period T vs semi-major axis a (average of aphelion and perihelion) of some Solar System orbits (crosses denoting Kepler's values) showing that a^3/T^2 is constant (green line)

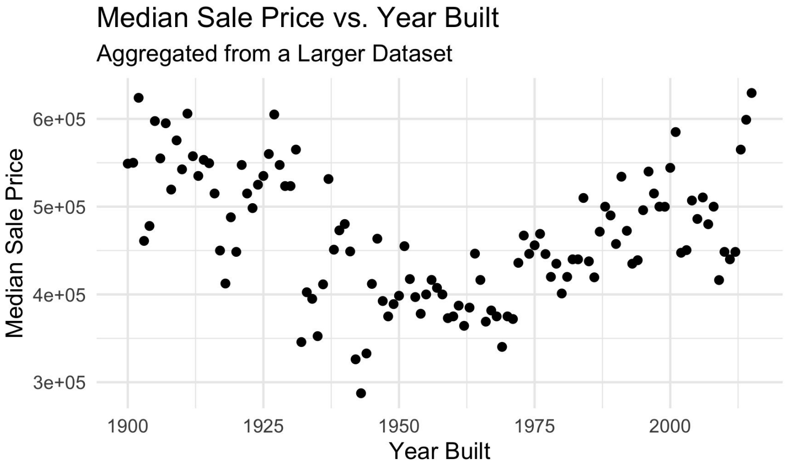
Models

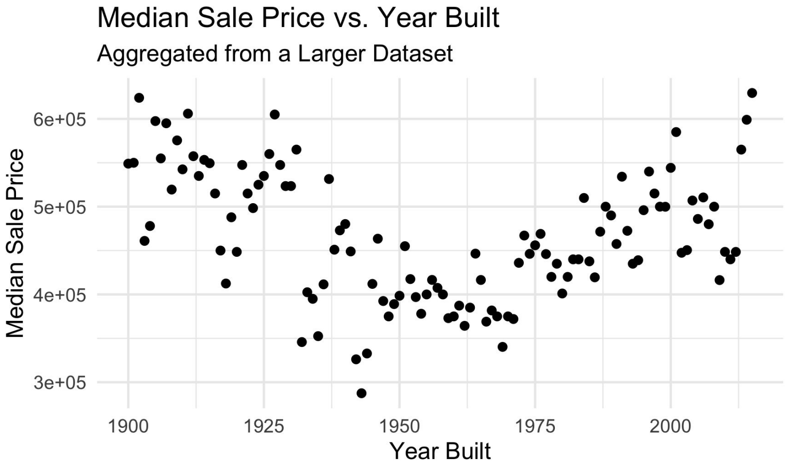
- Probabilistic models are those which seek to describe how random processes evolve.
  - Example: Global Forecasting Model (GFS), which seeks to model the behavior of weather systems.
  - Example: Black-Scholes-Merton Model, which seeks to produce estimates of prices of European-style options (e.g. Financial Data)
- Another model is the **Simple Random Walk** model, which you learn about in PSTAT 160A
- As statisticians, we know that there is **uncertainty** present in most data.
  - This is why we deal predominantly with probabilistic models as opposed to deterministic models.
  - **Uncertainty Quantification** is a branch of statistics that primarily deals with classification and understanding of such uncertainty/error.

Why Model?

PREDICTION

INFERENCE

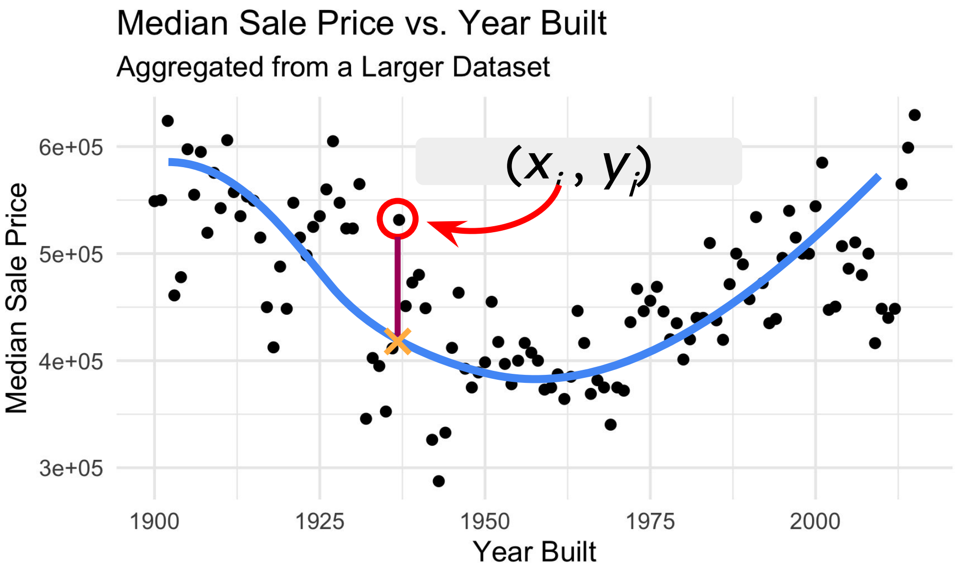




E.g. "What is the median price of a house in King County in 2026?"

E.g. "What is the nature of the relationship between time and the median price of King County Homes?"

Example Model



- **Model:** I can find the median housing price y_i at year x_i by taking a function of x_i and adding some random noise.

$$y_i = \boxed{f(x_i)} + \boxed{\varepsilon_i}$$

Explanatory Variable

Response variable

$$\boxed{\mathtt{y}} = \boxed{f}(\boxed{\mathtt{x}}) + \mathtt{noise}$$

Signal Function

"**Univariate**" = only one explanatory variable

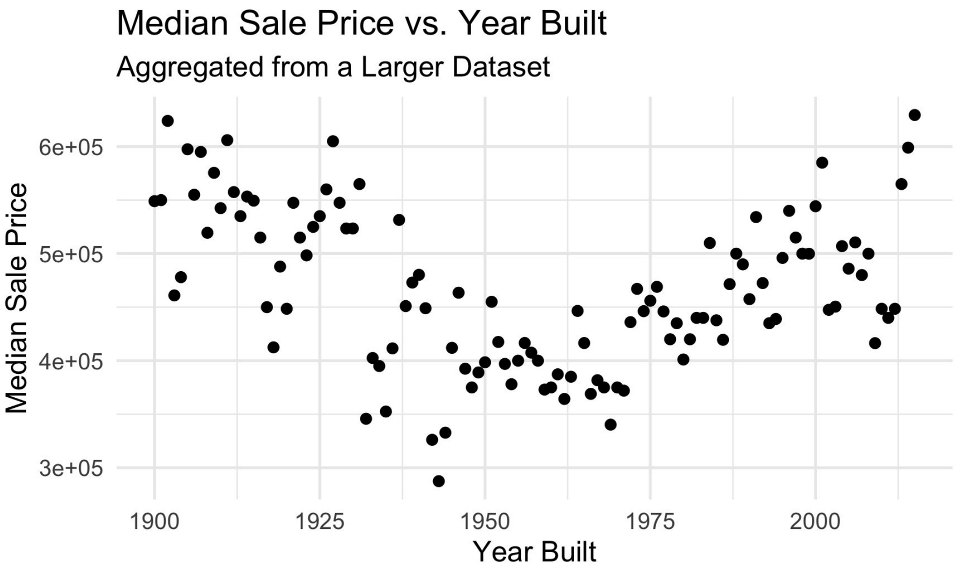
Types of Models

- There are two broad types of statistical models: **regression** and **classification**.
- Regression models arise when the response variable is numerical
  - E.g. predicting the average temperature on earth at a given time point
  - E.g. estimating the price of a stock tomorrow, given information about the company stock's past performance
- Classification models arise when the response variable is categorical
  - E.g. modeling the number of accidents on a freeway as a function of time
  - E.g. predicting whether a particular individual is predisposed to type II diabetes, based on various lifestyle factors
- You'll sometimes hear about "ANOVA" and "ANCOVA" models as well: ANOVA models arise when your predictor variable is categorical, and ANCOVA models arise when you have a mix of categorical and numerical predictor variables.

The Modeling Process

1. Propose a Model	*What type of model seems most appropriate?*
2. Choose a Loss Function	*How will we assess how well our model is fitting?*
3. Fit the Model	*What are the “best” parameter values, given the data we observed?*
4. Evaluate the Model	*How well are we doing?*

Housing Example: Propose a Model



- It seems plausible that housing prices are quadratically related to time:

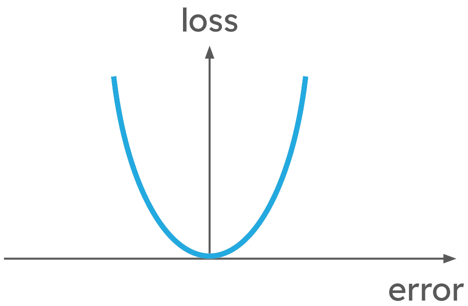
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

This is just the equation of a parabola!

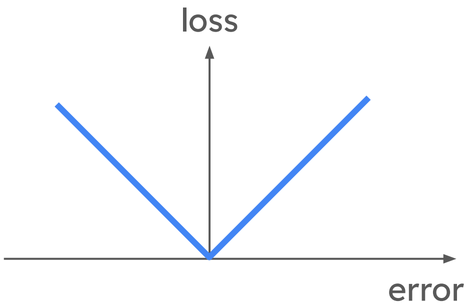
- By way, this is an example of **parametric modeling**, where we decompose our signal function into a series of *parameters* (in this case, the ß coefficients).
  - Contrast this to **nonparametric modeling**, which we may discuss at a later date.

Loss Functions

- In general, a **loss function** is a way to quantify how bad a prediction $\hat{y}$ is at predicting y.
  - If the prediction is good, we want low loss - if the prediction is bad, we want high loss.
- There are two commonly-used loss functions in regression:



Squared Loss: $L(y, \widehat{y}) = (y - \widehat{y})^2$



Absolute Loss: $L(y, \widehat{y}) = |y - \widehat{y}|$

- Squared-loss tends to be more highly influenced by outliers than absolute loss. This is something to keep in mind when selecting which loss function to use in a particular situation.

Housing Example: Loss Function

- We can aggregate losses across a dataset to obtain the **average loss** (sometimes called the **risk**):

$$R(\beta) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \widehat{y_i})$$

- For example, if we adopt a squared loss in the housing dataset, then our average loss would be

$$R(\beta_0, \beta_1, \beta_2) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

- So, given a candidate set of (β_0, β_1, β_2) the above quantity tells us how well/poorly our model is doing at fitting the data.

Housing Example: The Modeling Process So Far

1. Propose a Model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

2. Choose a Loss Function

We're using squared-loss

$$R(\beta_0, \beta_1, \beta_2) = \frac{1}{n}\sum_{i=1}^{n} \boxed{(y_i - \boxed{\beta_0 - \beta_1 x_i - \beta_2 x_i^2})^2}$$

We're fitting a parabola

3. Fit the Model

What are the "best" parameter values, given the data we observed?

4. Evaluate the Model

How well are we doing?

Housing Example: Estimate Parameters

- The “best” model will be the one that uses parameter values that minimize the risk (remember: low risk means our model is fitting the data well)
- Hence, we obtain estimators for the parameters by solving the following minimization problem:

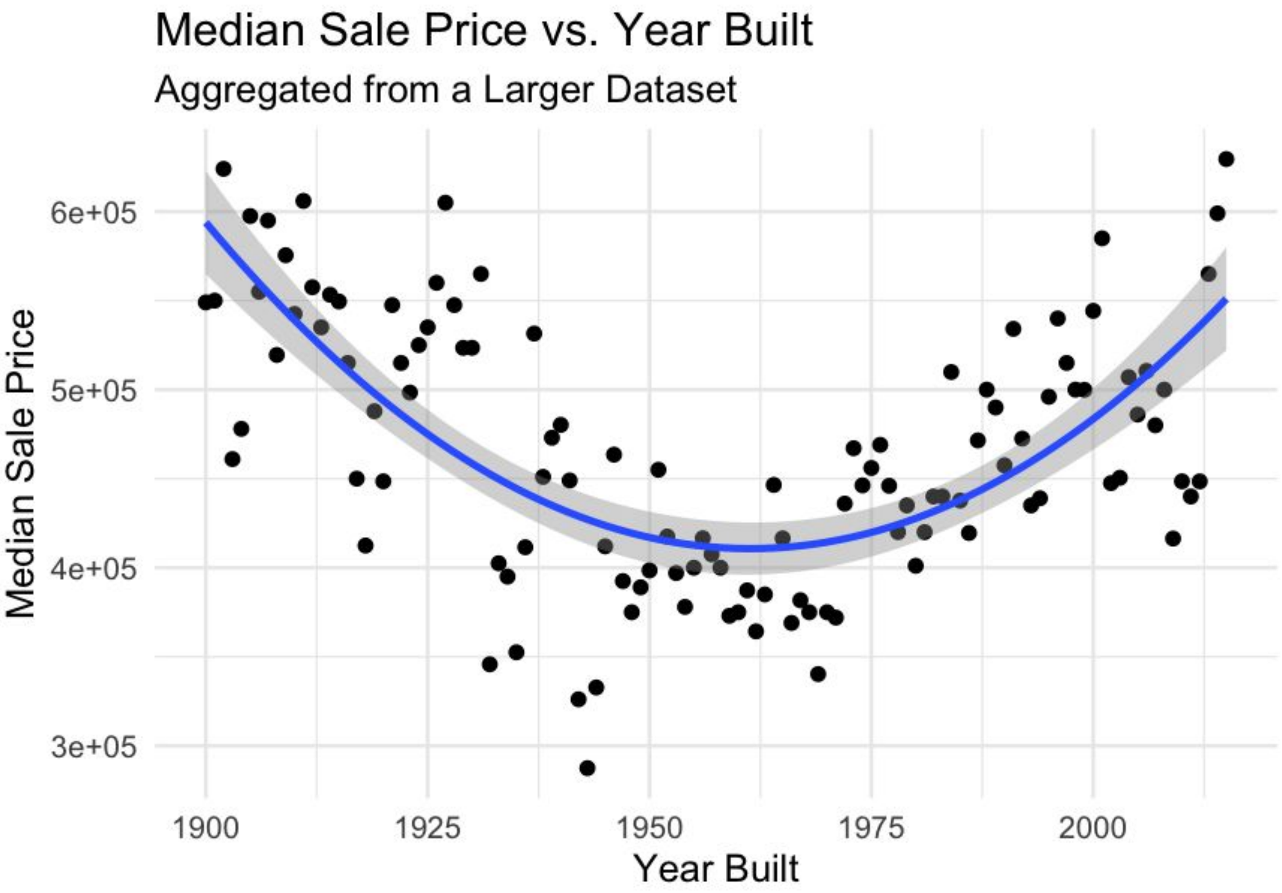
$$\widehat{\beta} = \boxed{\arg \min_{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(y_i, \widehat{y_i}) \right\}$$

This just means “find the β that minimizes…”

CAUTION

This step can be very computationally intensive, and there is current (active) research on making it more efficient.

Housing Example: Estimate Parameters



$$\widehat{\beta}_0 = 466035$$

$$\widehat{\beta}_1 = -134529$$

$$\widehat{\beta}_2 = 525909$$

Housing Example: Evaluate the Fit

- There are a few different ways we can evaluate the fit of our model:

1. **Visualizations:** if possible, plot the original data and overlay the fitted signal atop the scatterplot.
    a. Potential downsides: not always possible (especially in the case of multivariate data); not always possible to visually assess goodness-of-fit
2. **Performance Metrics:** e.g. root-mean-square (square root of risk under squared-error loss)
    a. Other metrics exist as well, depending on the context of your problem (e.g. Akaike Information Criterion, Bayesian Information Criterion for model selection)

Beyond Modeling

Machine Learning

- Statistical modeling can be thought of as a subset of **machine learning**.
- Machine Learning is divided into two main categories: **supervised learning** and **unsupervised learning**.
- In supervised learning, we have a response variable along with a collection of explanatory variables, and seek to model the relationship between the response and the explanatory variables.
  - This is more or less what we've talked about up until now!
- In unsupervised learning, we do not have a response variable; only a collection of explanatory variables.
  - The goal of machine learning in an unsupervised setting is to uncover patterns from the data.
  - E.g. given customer information from a store, we may seek to identify which types of customers are similar to each other (this is an example of **clustering**, the unsupervised analog of a classification problem).

Limitations of Modeling

- It’s a common misconception that data analysis is synonymous with modeling
    - There are a lot of very legitimate data analysis projects you can do without any models! (E.g. unsupervised learning; visualizations; etc.)
- If you do choose to use a model, though, there are some things to keep in mind (that we will discuss in a future workshop):
    - **Not all models generalize**. We fit models to data in the hopes that they help us say something about the world, but there are limitations to every model - just because you have a model that fits your data very well doesn’t mean you have a good model.
    - **Be mindful of the type of your model.** Don’t try to use squared error loss if you have categorical data; in fact, don’t even try to model categorical data yet (we’ll cover this in a future workshop)

Works Cited / Further Resources

- *An Introduction to Statistical Learning with Applications in R*, by James, Witten, Hastie, and Tibshirani
- *Hands-On Machine Learning with Scikit-Learn & TensorFlow*, by Géron
- UC Berkeley's Data 100 Course
- PSTAT 100 (link to an iteration I taught in Spring 2024)
- *Introduction to Modern Statistics*, by Çetinkaya-Rundel and Hardin

Let’s Get Coding!

https://bit.ly/dscol-mod1