

Some days you're the dog, some
days you're the hydrant

THOUGHT OF THE DAY

Objectives & Reminders

Objectives for today

1. *Mednet Case*
2. *Big Data - Market Basket (Part 1)*
3. *Kickstarter.com website*

Reminders:

- ⦿ Market basket – individual submission
- ⦿ Kickstarter (5) – group submission
- ⦿
- ⦿ Lab Feb 6
- ⦿ Lab Feb 13
- ⦿ Market Basket due by Feb 14th @ 23:55
- ⦿ Midterm – Feb 27th @ 6:30
- ⦿ Lab Mar 6
- ⦿ Lab Mar 13
- ⦿ Quiz Mar 31
- ⦿ FINAL project Update due by Mar 6th @ 23:55
- ⦿ FINAL Project due by Apr 7 @ 23:55
- ⦿ Final Presentations Apr 7 & 9
- ⦿ FINAL exam - TBD

AN OVERVIEW OF DATA MINING TECHNIQUES

Your task:

To understand the usage of different types of business intelligence tools

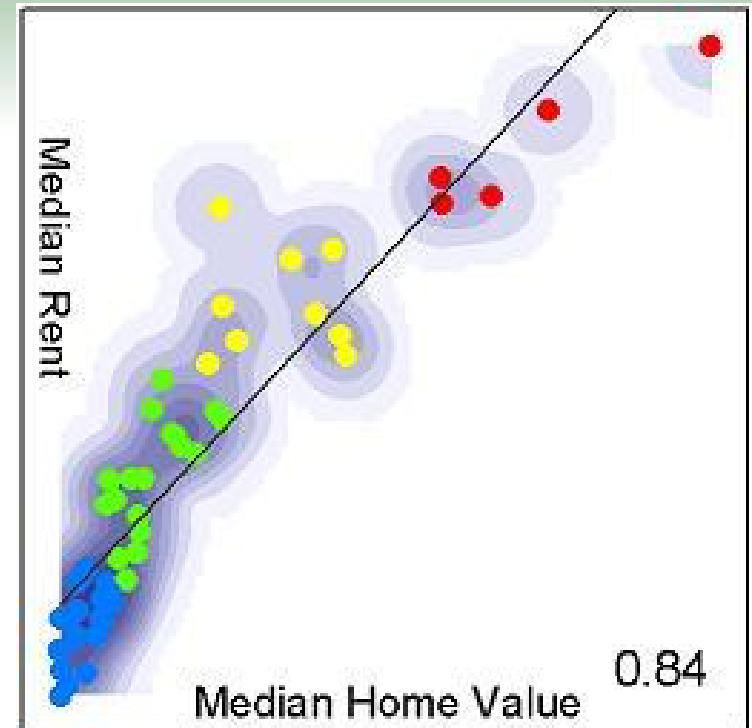
DATA MINING TASKS

- Four classes of data mining tasks
 - Unsupervised
 - ✓ Clustering
 - ✓ Association Detection
 - Supervised
 - ✓ Classifications
 - ✓ Regressions



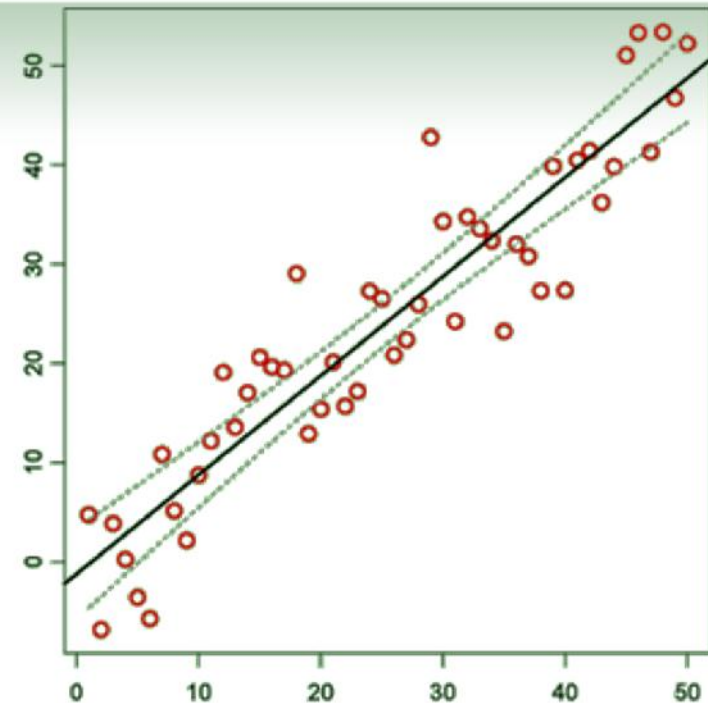
UNSUPERVISED DATA MINING

- ⊙ Analysts do not create model before running analysis . Explore the data to find some intrinsic structures in them
- ⊙ Apply data-mining technique and observe results
- ⊙ Hypotheses created after analysis as explanation for results
- ⊙ The goal is not known or specified and task is to determine the best groupings
- ⊙ Example: cluster analysis, association rules



SUPERVISED DATA MINING

- Model developed before analysis (more for predictive analysis)
- Statistical techniques used to estimate parameters
- Examples:
 - Regression analysis
 - Neural networks
 - Decision Trees
- The goal is known or specified and the task is to determine relationship among independent variables and dependent variable (goal)



Regression can help determine the relationship between variables

$$\text{Performance} = \beta_1 + \beta_2 \times \text{skill level}$$

Example of Research paper using Supervised data mining

103/6/1632.full

Analysis of meal patterns with the use of supervised data mining techniques—artificial neural networks and decision trees 1'2'3

Áine P Hearty and Michael J Gibney

[+ Author Affiliations](#)

Abstract

Background: At present, the analysis of dietary patterns is based on the intake of individual foods. This article demonstrates how a coding system at the meal level might be analyzed by using data mining techniques.

Objective: The objective was to evaluate the usability of supervised data mining methods to predict an aspect of dietary quality based on dietary intake with a food-based coding system and a novel meal-based coding system.

Design: Food consumption databases from the North-South Ireland Food Consumption Survey 1997–1999 were used. This was a randomized cross-sectional study of 7-d recorded food and nutrient intakes of a representative sample of 1379 Irish adults. Meal definitions were recorded by the respondent. A healthy eating index (HEI) score was developed. Artificial neural networks (ANNs) and decision trees were used to predict quintiles of the HEI based on combinations of foods consumed at breakfast and main meals.

Results: This study applied both data mining techniques to the food and meal-based coding systems. The ANN had a slightly higher accuracy than did the decision tree in relation to its ability to predict HEI quintiles 1 and 5 based on the food coding system (78.7% compared with 76.9% and 71.9% compared with 70.1%, respectively). However, the decision tree had higher accuracies than did the ANN on the basis of the meal coding system (67.5% compared with 54.6% and 75.1% compared with 72.4%, respectively).

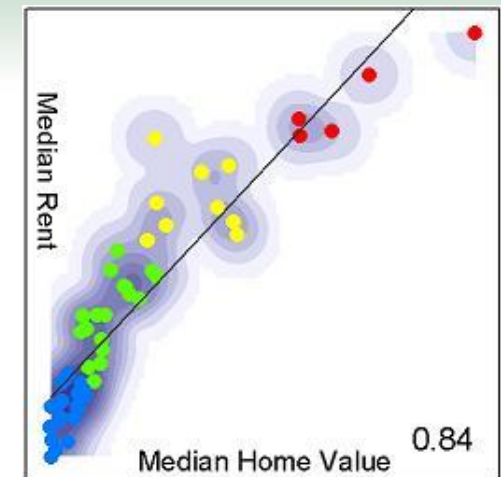
Conclusions: ANNs and decision trees were successfully used to predict an aspect of dietary quality. However, further exploration of the use of ANNs and decision trees in dietary pattern analysis is warranted.

CLUSTERING

(UNSUPERVISED)

Cluster Analysis

- Similar records (or characteristics) are grouped together
- Does not rely on predefined categories*** - being grouped together on the basis of self-similarity



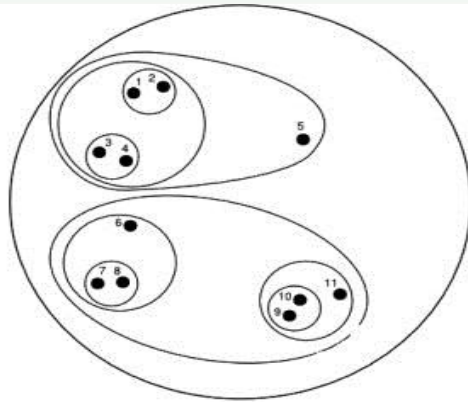
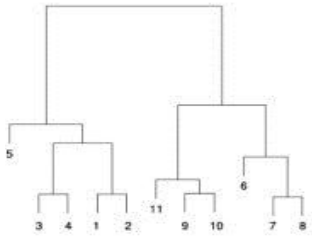
Example

- Market segmentation: Identify customers with similar buying behavior

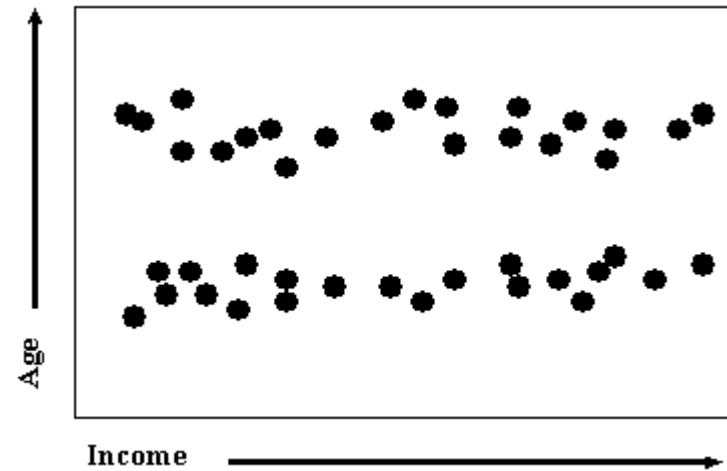
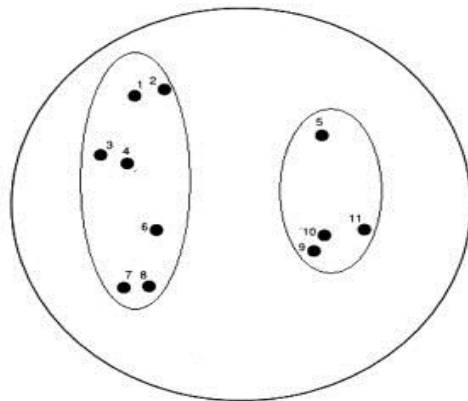
Common algorithm: K-means Analysis

CLUSTERING

(a)



(b)



CLASSIFICATION

(SUPERVISED)

- **Classification**
 - Arrange the data into predefined groups
 - Difference between Clustering and Classification
 - Depends on whether categories are **predefined** or not
- Examples
 - Classify credit applicants as low, medium, or high risk
 - Classify customers as “loyal” vs. “likely to terminate contract”
 - Classify emails as legitimate or spam
- Common algorithms:
 - Decision Tree Analysis

ASSOCIATION RULES

(UNSUPERVISED)

◎ Association rules

- ◎ Determine which behaviors/outcomes go together
- ◎ Find relationships among attributes in data that frequently occur together

◎ Examples

- ◎ **Market basket analysis:** *determine what things go together in a shopping cart at the supermarket.*



MARKET-BASKET ANALYSIS

Your Task:

To know how to conduct market-basket analysis

MARKET-BASKET ANALYSIS

- ◎ Data-mining technique for determining sales patterns
- ◎ Shows products that customers tend to buy **together**
- ◎ Example:

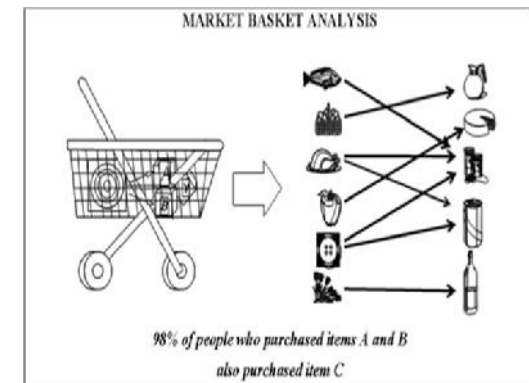
On Thursday nights, people who buy diapers may also buy beer



MARKET-BASKET ANALYSIS

Product affinities → **cross-selling** opportunities

Product		Association		Lift	Confidence
Orblt Sleeping Pad		Orblt Stuff Sack		222	37%
Bambini Tights Children's		Ramhini Crewneck Sweater Children's		195	52%
Silk Crew Women's		Silk Long Johns Women's		304	73%
Cascade Entrant Overmitts		Polartec 300 Double Mitts		51	48%



Item affinity – defines the likelihood of two (or more) items being sold together

Market-Basket Analysis

◎ **Support:** Probability that certain product(s) is(are) bought (it is the percentage of transactions that contain both A and B)

◎ $P(A), P(A \& B)$

◎ **Support Count:** Number of times that certain product(s) has(have) been bought

◎ $\sigma(A), \sigma(A \& B)$

$$\text{Support} = \frac{\text{Support count}}{\text{Number of transactions}}$$



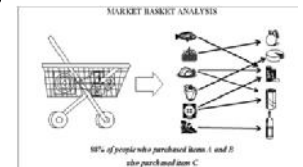
Market-Basket Analysis

- ◎ **Confidence**: a *conditional probability* - given a person bought product A, the likelihood he/she will also buy product B

$$P(B|A) = \frac{P(A \& B)}{P(A)} = \frac{\frac{\sigma(A \& B)}{\text{No. transactions}}}{\frac{\sigma(A)}{\text{No. transactions}}} = \frac{\sigma(A \& B)}{\sigma(A)}$$

- ◎ **Lift**: the ratio of confidence to the base probability of buying an item

$$\frac{P(B|A)}{P(B)}$$



Lift tells you how much better than chance item x will appear in the cart if you already know that item Y is in the cart.

Market-Basket Analysis

◎ Binary Representation

Items

Transactions {

Tran #	Milk	Diaper	Beer
1	1	0	0
2	0	1	1
3	1	1	1
4	1	1	1
5	1	1	0

◎ Association Rules:

$$\underbrace{\{\text{Milk, Diaper}\}}_{\text{Antecedent}} \rightarrow \underbrace{\{\text{Beer}\}}_{\text{Consequent}}$$

Market-Basket Analysis

- For an Association Rule:

$\{\text{Item1}, \text{Item2}, \dots\} \rightarrow \{\text{Item3}, \text{Item4}, \dots\}$

{Antecedent} → {Consequent}

- Confidence Level:**

$$P(\text{Consequent} \mid \text{Antecedent}) = \frac{P(\text{Antecedent} \ \& \ \text{Consequent})}{P(\text{Antecedent})}$$

- Lift Ratio:**

$$\text{Lift} = \frac{P(\text{Consequent} \mid \text{Antecedent})}{P(\text{Consequent})}$$



EXERCISE

Tran #	Milk	Diaper	Beer
1	1	0	0
2	0	1	1
3	1	1	1
4	1	1	1
5	1	1	0

◎ **Rule: {Milk, Diaper} \rightarrow {Beer}**

◎ **Support:**

◎ $\sigma(\text{Milk \& Diaper}) = 3$ $P(\text{Milk \& Diaper}) = 3/5 = .6$

◎ $\sigma(\text{Beer}) = 3$ $P(\text{Beer}) = 3/5 = .6$

◎ $\sigma(\text{M\&D\&B}) = 2$ $P(\text{M\&D\&B}) = 2/5 = .4$

EXERCISE

Tran #	Milk	Diaper	Beer
1	1	0	0
2	0	1	1
3	1	1	1
4	1	1	1
5	1	1	0

⊙ **Rule: {Milk, Diaper} → {Beer}**

⊙ **Confidence:**

$$\begin{aligned}
 \odot \quad P(\text{Beer} \mid \text{Milk\&Diaper}) &= P(\text{B\&M\&D}) / P(\text{M\&D}) \\
 &= \sigma(\text{M\&D\&B}) / \sigma(\text{M\&D}) \\
 &= .4 / .6 = 2 / 3 = .667
 \end{aligned}$$

EXERCISE

Tran #	Milk	Diaper	Beer
1	1	0	0
2	0	1	1
3	1	1	1
4	1	1	1
5	1	1	0

⊙ **Rule: {Milk, Diaper} → {Beer}**

⊙ **Lift (econometrics = Lorenz or power curve)**

$$P(B|A)$$

$$P(B)$$

⊙ Lift considers confidence of the rule and the overall data set

⊙ **Lift of 1 – probability of occurrence of antecedent and consequent are independent of each other.**

When lift > 1 then the rule is better at predicting the result than guessing. When lift < 1, the rule is doing worse than informed guessing

⊙ $P(\text{Beer} \mid \text{Milk\&Diaper}) / P(\text{Beer}) = .667 / .6 = 1.11$

ASSOCIATION RULES

- ◎ Need to find **meaningful** association rules
 - ◎ **Minimum support requirement**
 - Supports need to be large enough to be *statistically significant*
 - ◎ **Minimum confidence requirement**
 - Higher confidence indicates stronger correlation
 - ◎ **The lift ratio needs to be high enough**



EVALUATE ASSOCIATION RULES

- ⊙ Is Rule {Milk, Diaper} → {Beer} a “good association rule?”
- ⊙ Assume a good association rule needs to satisfy the following requirements
 - ⊙ minimum support requirement as 0.1 (Probability that certain product(s) is(are) bought)
 - ⊙ minimum confidence requirement 0.6 (*conditional probability* - given a person bought product A, the likelihood he/she will also buy product B)
 - ⊙ lift ratio greater than 1 (Lift considers confidence of the rule and the overall data set)
- ⊙ Yes. All three conditions are satisfied
 - ⊙ $P(M\&D\&B) = 2/5 = .4$
 - ⊙ $P(Beer \mid Milk\&Diaper) = .667$
 - ⊙ $P(Beer \mid Milk\&Diaper) / P(Beer) = .667 / .6 = 1.11$

Market Basket Analysis

- **LIMITATIONS**

- takes over 18 months to implement
- market basket analysis only identifies hypotheses, which need to be tested
- measurement of impact needed
- difficult to identify product groupings
- complexity grows exponentially

MARKET BASKET ANALYSIS

- **BENEFITS:**

- simple computations
- can be undirected (don't have to have hypotheses before analysis)
- different data forms can be analyzed

WHAT IS THE CONFIDENCE OF THE ASSOCIATION RULE $\{XBOX, GAME1\} \rightarrow \{GAME2\}$?

Tran #	Xbox	Game1	Game2
1	1	0	0
2	0	0	1
3	1	1	1
4	1	1	0
5	1	1	1

1. $3/3=1$

2. $2/3=0.67$

3. $3/5=0.6$

4. $3/2=1.5$

Homework

MBA - POINTS TO CONSIDER

- In MBA, the # of each items bought is not relevant
 - Example: buying one textbook vs 3 textbooks
- Only transactions of more than one item is considered as data
- Input data must be clean and error free.

[Video-example](#)

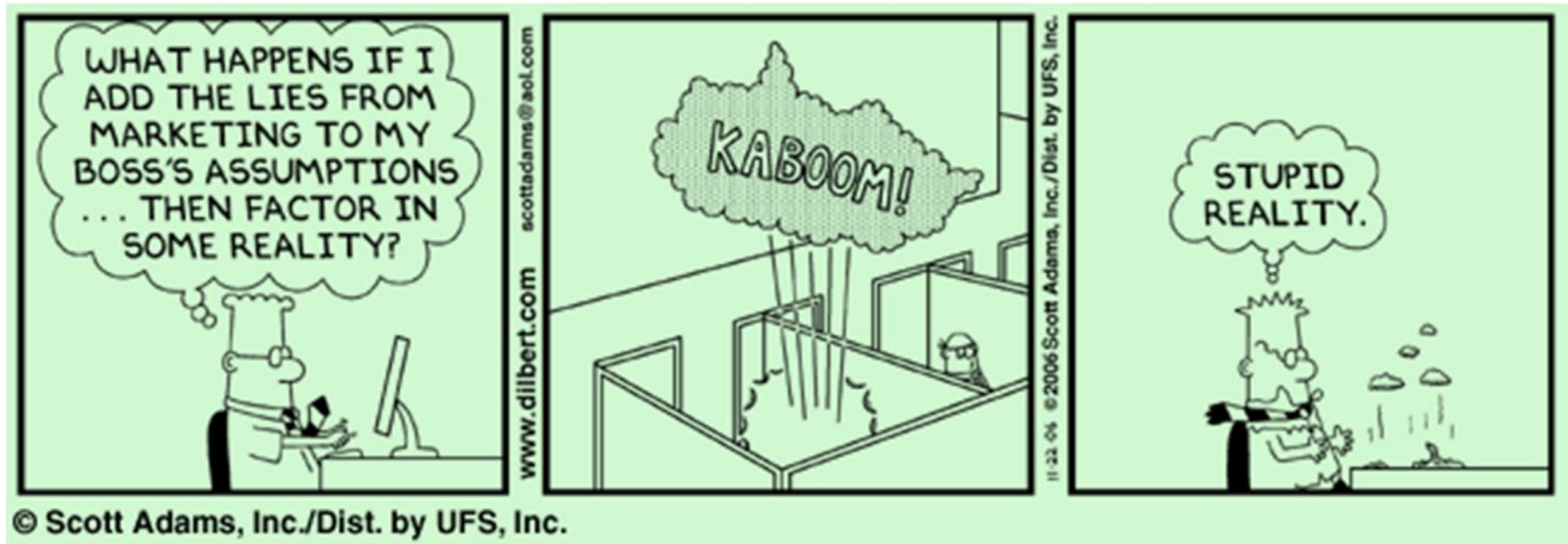
DATA MINING LINKAGES - SUPERVISED AND UNSUPERVISED



- ③ When lift is _____ then the rule is better at predicting the result than guessing. When lift _____, the rule is doing worse than informed guessing
- ③ **>1, < 1**

- ⊙ T or F – Market Basket is Unsupervised Data Mining Technique

DILBERT...



🎯 KICKSTARTER.COM

SUPERVISED

- Optional for students

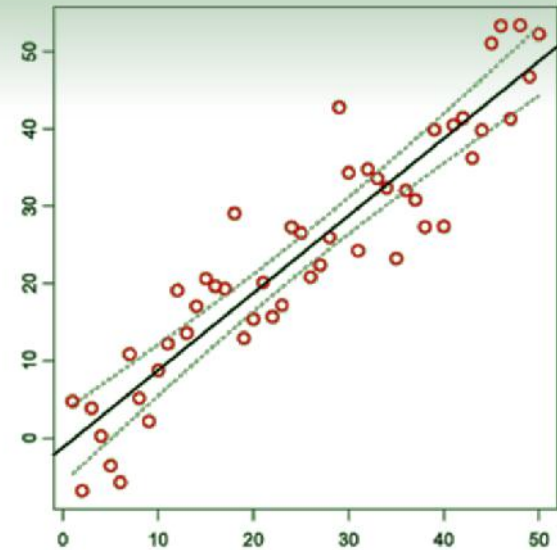
REGRESSION

Regression

- Attempt to find a **function** which models the data with the least error
- Determine the relationship between some unknown variable and a set of known variables
- Involves use of some training data to determine parameters

Example – estimate β_1 and β_2

$$\text{Performance} = \beta_1 + \beta_2 \times \text{skill level}$$



Regression can help determine the relationship between variables

EXAMPLE: UNDERSTANDING HOUSE PRICES

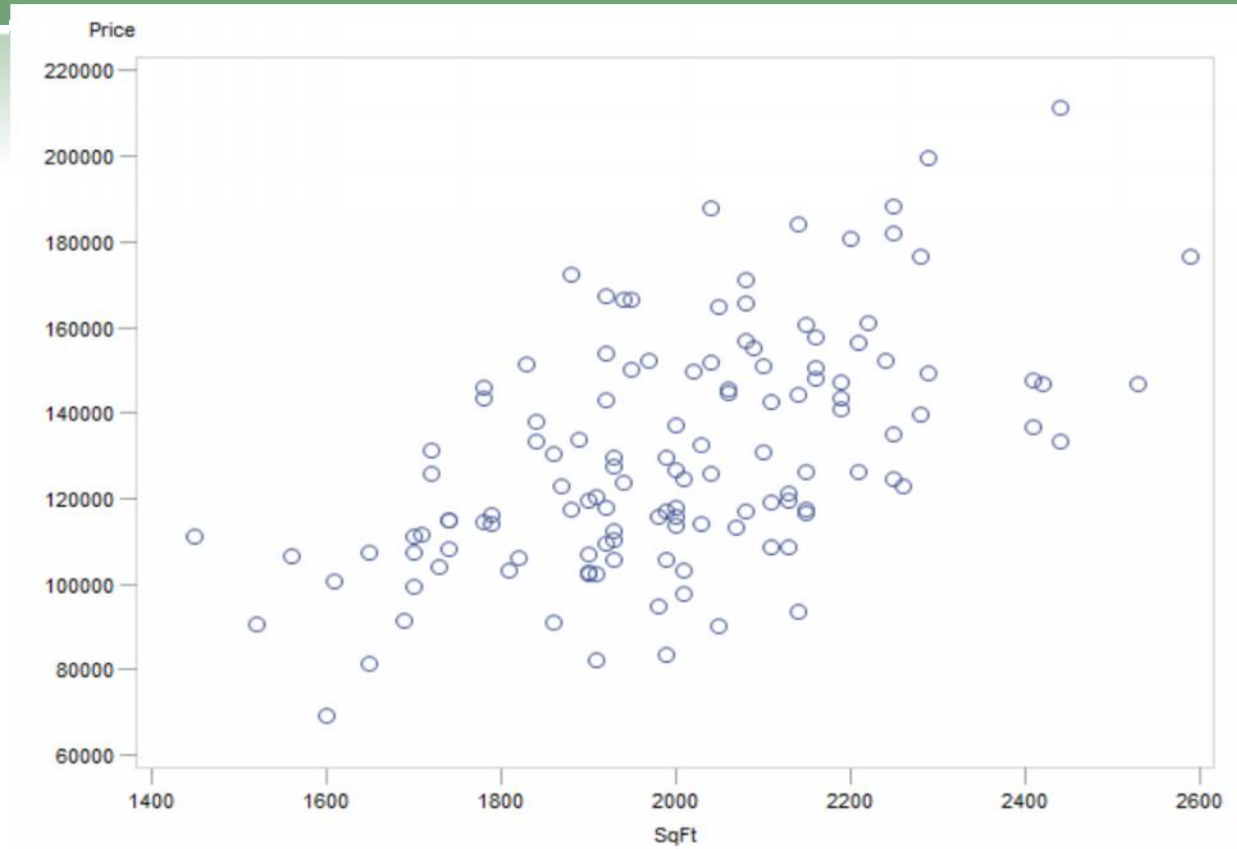
HomeID	Price	SqFt	Bedrooms	Bathroom	Offers	Brick	Neighborhood
1	114300	1790	2	2	2	No	East
2	114200	2030	4	2	3	No	East
3	114800	1740	3	2	1	No	East
4	94700	1980	3	2	3	No	East
5	119800	2130	3	3	3	No	East
6	114600	1780	3	2	2	No	North
7	151600	1830	3	3	3	Yes	West
8	150700	2160	4	2	2	No	West
9	119200	2110	4	2	3	No	East

⊙ Available data on homes sold in a metro area

- ⊙ price fetched
- ⊙ square footage
- ⊙ # of bedrooms and bathrooms
- ⊙ # of offers it received
- ⊙ whether its exteriors is brick
- ⊙ neighborhood it is located in

⊙ Goal: Understand why some houses sell for more than others!

SCATTERPLOT



- Scatterplot shows a relationship between sqft and price
- How would you best characterize this relationship?

EXAMPLE: UNDERSTANDING HOUSE PRICES

HomeID	Price	SqFt	Bedrooms	Bathroom	Offers	Brick	Neighborhood
1	114300	1790	2	2	2	No	East
2	114200	2030	4	2	3	No	East
3	114800	1740	3	2	1	No	East
4	94700	1980	3	2	3	No	East
5	119800	2130	3	3	3	No	East
6	114600	1780	3	2	2	No	North
7	151600	1830	3	3	3	Yes	West
8	150700	2160	4	2	2	No	West
9	119200	2110	4	2	3	No	East

- ⊙ Exploring data: why some houses sell for more than others
- ⊙ Goal: precisely quantify **how much more** a house sells with e.g. an additional bathroom or 200 extra sqft
 - ⊙ Which variables carry more information than others
 - ⊙ Which carry no practical information at all
- ⊙ Q: Can we get this in scatterplot/correlation?

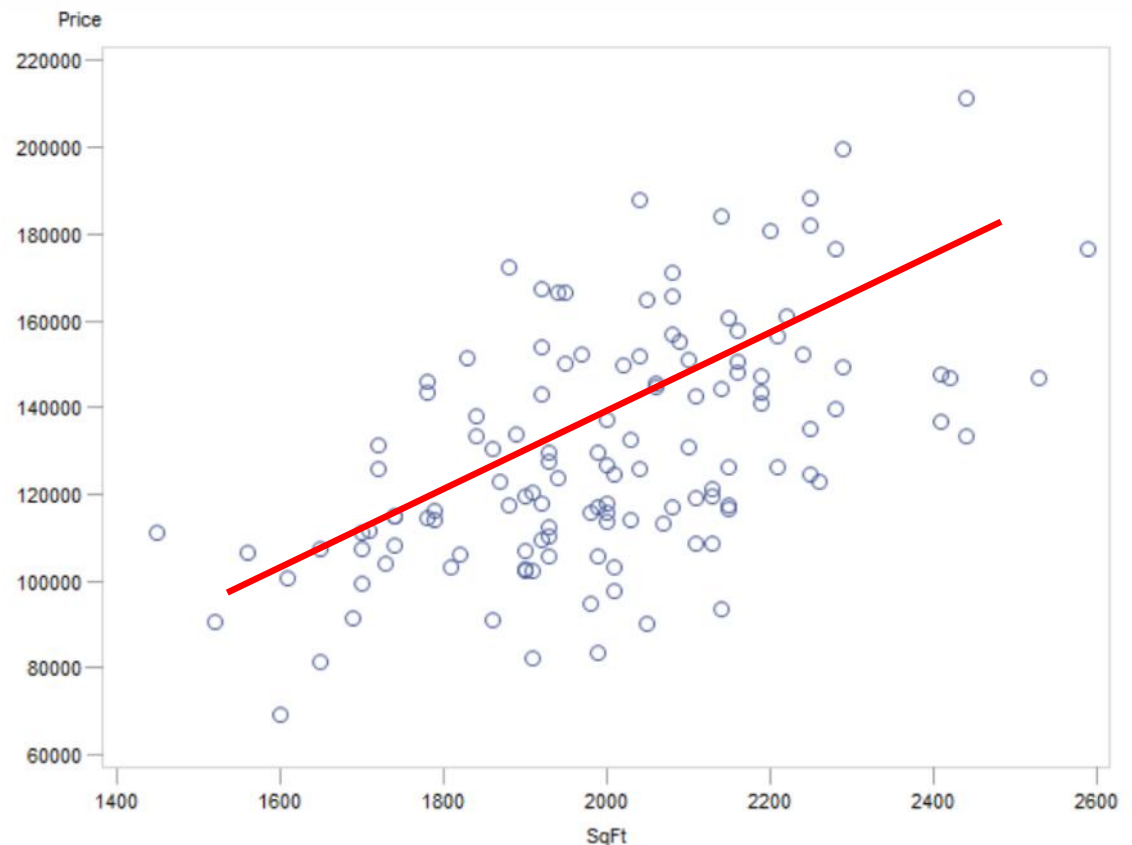
FUNDAMENTAL ASSUMPTION OF LINEAR REGRESSION

◎ Linearity

- ◎ If a linear relationship (red line) is a sufficient approximation to our data, we will have a mathematical equation

$$\text{Price} = a + b \text{ SqFt}$$

- ◎ How to estimate coefficients (a and b)?



USING REGRESSION FOR PREDICTION

- ⊙ Price = $-\$10,091 + \$70 * \text{SqFt}$
- ⊙ What is the predicted price of a house with 2,000 sqft?

Usage of Regression Models

- ⊙ We use regression models to
 - ⊙ Obtain predictions
 - Predictions about future sales, given a certain amount of advt expenditures
 - ⊙ e.g., how much advt do I need to reach a certain sales target?
 - ⊙ Obtain insight into the economic relationships
 - Insight into how exactly advt expenditures and sales relate
 - ⊙ e.g., is advt an effective driver of sales?