SQL for Analytics
Data Summaries

IOE 373 Lecture 07

Topics

- Analytics Concepts/Process
- Frequency/Distribution Tables
- IIF Function
- Format Function

Other Useful Queries (For Analytics)

- Data Analytics requires data processing to prepare data analysis table.
  - Extract or calculate predictor variables (input or independent variables, X_i) and target variable (output or dependent variables, Y)

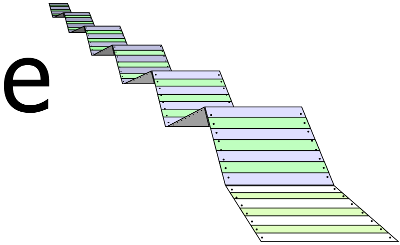
$$Y=f(X_i)$$

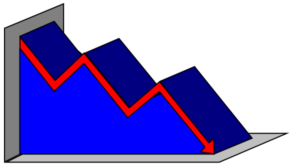
These are typically empirical models, meaning that we use data to “fit” the function or estimate the parameters that help us predict the target based on the inputs

Data Analytics?

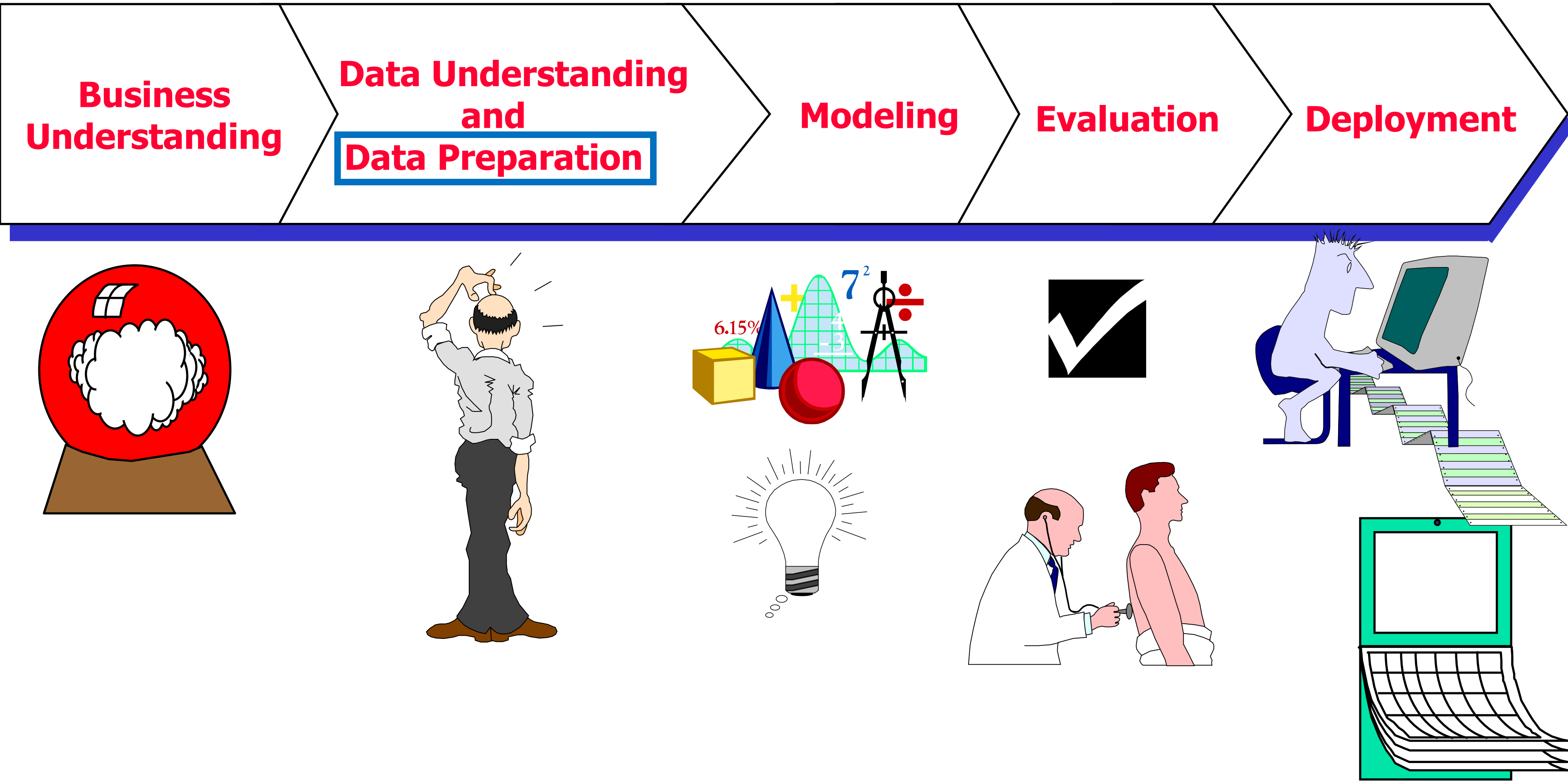
- One of many definitions:

The process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems





Modeling Process



Business Understanding

Data Understanding and Data Preparation

Modeling

Evaluation

Deployment

Some “Classic” Examples

- Customer Segmentation Models
- Retention Models (Loyalty or Churn)
- Customer Acquisition Models
- Campaign Management Models
- Cross-selling and Up-selling
- Forensic Analysis Models (Fraud detection)
- Pricing Models
- Customer Value/Profitability Models
- Market Basket/Association Analysis

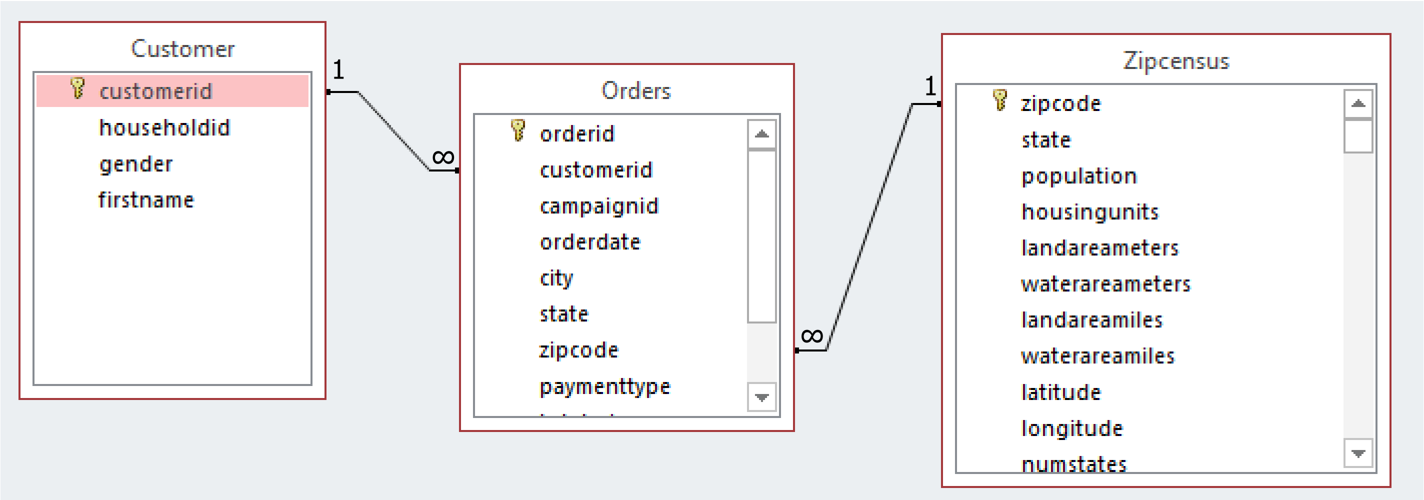
Most Common Modeling Methods

- Regression - A linear equation of predictor variables.
- Logistic Regression – A variation of linear regression used to predict probabilities
- Tree Methods (CART, Random Forests) - A hierarchical structure of significant variables in order of importance
- Neural Networks – Multiple types of methods that assign weights to input variables in a non-linear fashion.

Case Study Example: Marketing Campaign

- Analyze database to obtain general summaries and statistics about households
    - Customers per household
    - Frequency analysis of households
    - Summaries by payment type
- Data Analysis Table for predictive model(s)
    - Get predictors or input variables
    - Get response(s) or target variables
    - Predictive model to estimate/calculate response based on the predictors or inputs

Purchases Database



Customer
- customerid
- householdid
- gender
- firstname

Orders
- orderid
- customerid
- campaignid
- orderdate
- city
- state
- zipcode
- paymenttype

Zipcensus
- zipcode
- state
- population
- housingunits
- landareameters
- waterareameters
- landareamiles
- waterareamiles
- latitude
- longitude
- numstates

Customers in Household

```
SELECT householdid, COUNT(*) as
numinhousehold
        FROM customer
        GROUP BY householdid
        HAVING COUNT(*) <=10
```

Distribution of Customers Per Household

- - How many households have 1 customer, 2 customers, 3 customers, etc…?
  - For each possible number of people per household we need the frequency (e.g. count of households that have 1 person, 2 persons, etc.)
  - We already have a table listing the number of people per household (previous query):

```
SELECT householdid, COUNT(*) as numinhousehold
     FROM customer
     GROUP BY householdid
```

Distribution of Customers Per Household

- Let’s get a table showing the numinhousehold and corresponding count….and use the previous query as the source of the distribution table…

```
SELECT numinhousehold, COUNT(*) as numhh
FROM ?
```

Distribution of Customers Per Household

- Let’s get a table showing the numinhousehold and corresponding count….and use the previous query as the source of the distribution table…

```
SELECT numinhousehold, COUNT(*) as numhh
FROM (SELECT householdid, COUNT(*) as
numinhousehold
      FROM customer
      GROUP BY householdid)
GROUP BY numinhousehold
ORDER BY numinhousehold
```

What if I want to filter out households with more than 10 people?

Distribution of Customers Per Household – Filter out households over 10 people

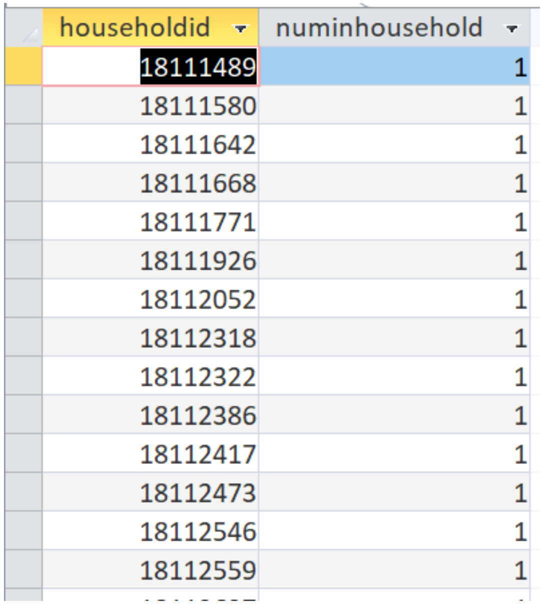
```
SELECT numinhousehold, COUNT(*) as numhh,
format(COUNT(*)/(select count(*) from
customer),"Percent") AS PercentNum
FROM (SELECT householdid, COUNT(*) as
numinhousehold
      FROM customer
      GROUP BY householdid Having
Count(*)<=10)
GROUP BY numinhousehold
ORDER BY numinhousehold
```

Complete Query

Dist_NumCustomers_InHouseHold	
numinhousehold	numhh
1	151987
2	1776
3	26
4	1

Source (or Subquery)

```
SELECT householdid, COUNT(*) as numinhousehold
      FROM customer
      GROUP BY householdid Having Count(*)<=10
```



householdid	numinhousehold
18111489	1
18111580	1
18111642	1
18111668	1
18111771	1
18111926	1
18112052	1
18112318	1
18112322	1
18112386	1
18112417	1
18112473	1
18112546	1
18112559	1

Summary by Payment Type

Get a summary by Payment Type where:

- We count the number of transactions per the following categories/buckets
  - Between 0 and $10
  - Between $10 and $100
  - Between $100 and $1,000
  - Over $1,000
  - Get a total revenue by Payment Type

IIF Function

- IIF (condition, truepart, falsepart)
- Can be nested within other functions, e.g.
  - SUM (IIf (expression, 1, 0)) would give you a count of records that meet a condition
  - This is equivalent to a COUNTIF in excel which is not existent in standard SQL…

Format Function

- Format (expression, [format])

Format	Explanation
General Number	Displays a number without thousand separators.
Currency	Displays thousand separators as well as two decimal places.
Fixed	Displays at least one digit to the left of the decimal place and two digits to the right of the decimal place.
Standard	Displays the thousand separators, at least one digit to the left of the decimal place, and two digits to the right of the decimal place.
Percent	Displays a percent value - that is, a number multiplied by 100 with a percent sign. Displays two digits to the right of the decimal place.
Scientific	Scientific notation.
Yes/No	Displays No if the number is 0. Displays Yes if the number is not 0.
True/False	Displays False if the number is 0. Displays True if the number is not 0.
On/Off	Displays Off if the number is 0. Displays On is the number is not 0.

https://msdn.microsoft.com/en-us/library/office/jj720239.aspx

Summary by Payment Type

paymenttype	cnt_0_10_USD	cnt_10_100USD	cnt_100_1000USD	cnt_1000USD	cnt	revenue	Percent Revenue
??	298	11	4	0	313	$1,184.17	0.01%
AE	1483	36093	9341	465	47382	$4,656,038.04	33.96%
DB	5356	6524	823	36	12739	$471,008.74	3.44%
MC	1862	38458	6797	201	47318	$3,302,579.39	24.09%
OC	3643	4042	518	11	8214	$264,647.83	1.93%
VI	3206	62993	10545	273	77017	$5,013,438.13	36.57%

Summary by Payment Type

```
SELECT paymenttype, SUM(IIF(0 <= totalprice AND
totalprice < 10, 1, 0)) AS cnt_0_10_USD, SUM(IIF(10
<= totalprice AND totalprice < 100,1,0)) AS
cnt_10_100USD, SUM(IIF(100 <= totalprice AND
totalprice < 1000,1,0)) AS cnt_100_1000USD,
SUM(IIF(totalprice >= 1000,1,0)) AS cnt_1000USD,
COUNT(*) AS cnt, format(SUM(totalprice),"CURRENCY")
AS revenue, format(SUM(totalprice)/(select
sum(totalprice) from orders),"percent") AS
PercentRevenue
FROM orders
GROUP BY paymenttype
ORDER BY paymenttype;
```

Other Summaries

- List of Max TotalPrice by State (eliminate nulls and unknown characters)
- Get the top 10 states based on Max Totalprice
  - Order from smallest to largest!
- Include the City name for where the Max TotalPrice by State comes from

- List of Max TotalPrice by State (eliminate nulls and unknown characters)

- List of Max TotalPrice by State (eliminate nulls and unknown characters)

```
SELECT State, max(Totalprice) as MaxTotalState
FROM Orders
WHERE State IS NOT NULL AND State <>"."
GROUP BY State
```

- Get the top 10 states based on Max Totalprice
  - Order from smallest to largest!

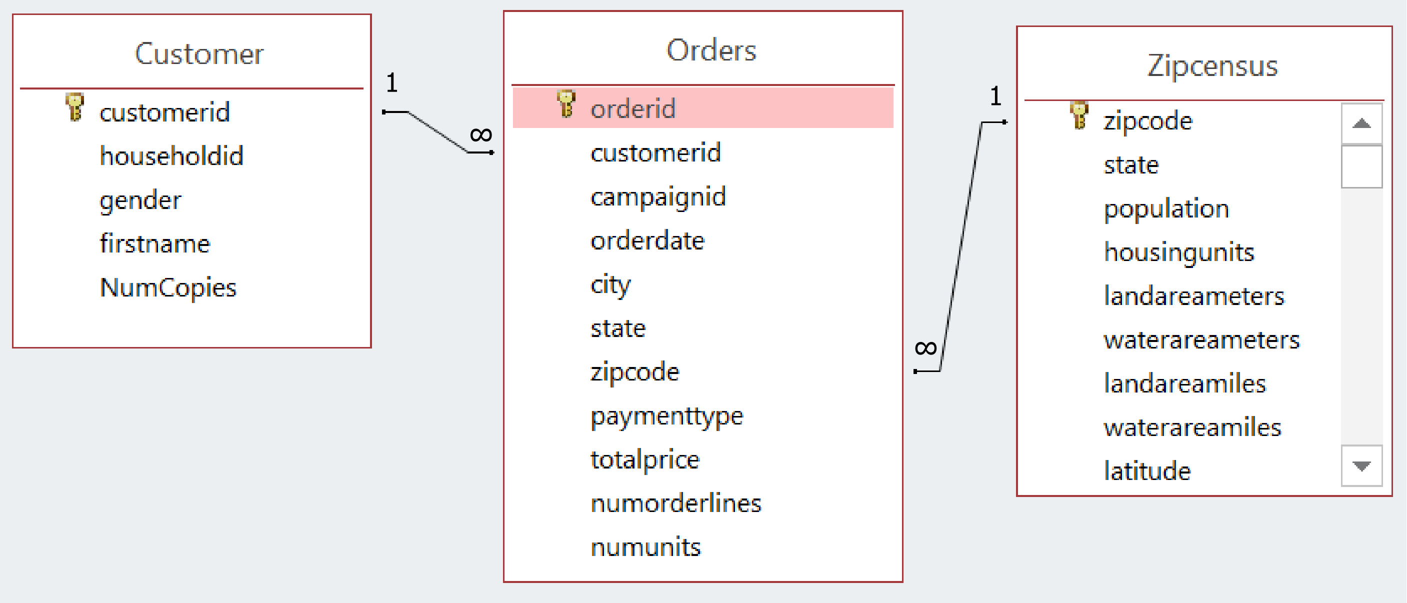
- Get the top 10 states based on Max Totalprice

```
SELECT TOP 10 State, format(max(Totalprice),
"#,###.00") as MaxTotalState
    FROM Orders
    WHERE State IS NOT NULL AND State <>"."
    GROUP BY State
    ORDER BY max(Totalprice) DESC
```

- Get the top 10 states based on Max Totalprice
  - Order from smallest to largest!

```
SELECT * FROM
   (SELECT TOP 10 State, format(max(Totalprice),
"#,###.00") as MaxTotalState
   FROM Orders
   WHERE State IS NOT NULL AND State <>"."
   GROUP BY State
   ORDER BY max(Totalprice) DESC)
ORDER BY MaxTotalState
```

- For each State, include the City name from where the Max TotalPrice comes from



Customer

customerid
householdid
gender
firstname
NumCopies

Orders

orderid
customerid
campaignid
orderdate
city
state
zipcode
paymenttype
totalprice
numorderlines
numunits

Zipcensus

zipcode
state
population
housingunits
landareameters
waterareameters
landareamiles
waterareamiles
latitude

- For each State, include the City name from where the Max TotalPrice comes from

```
SELECT City, State, Totalprice
FROM Orders AS x
WHERE State IS NOT NULL AND State <>"." AND
Totalprice=
  (Select max(Totalprice) FROM Orders as y
      WHERE x.State=y.State)
ORDER BY State
```