CS434 Notebook

April 19

2017

Data Mining and Data Warehouse

Table of Contents

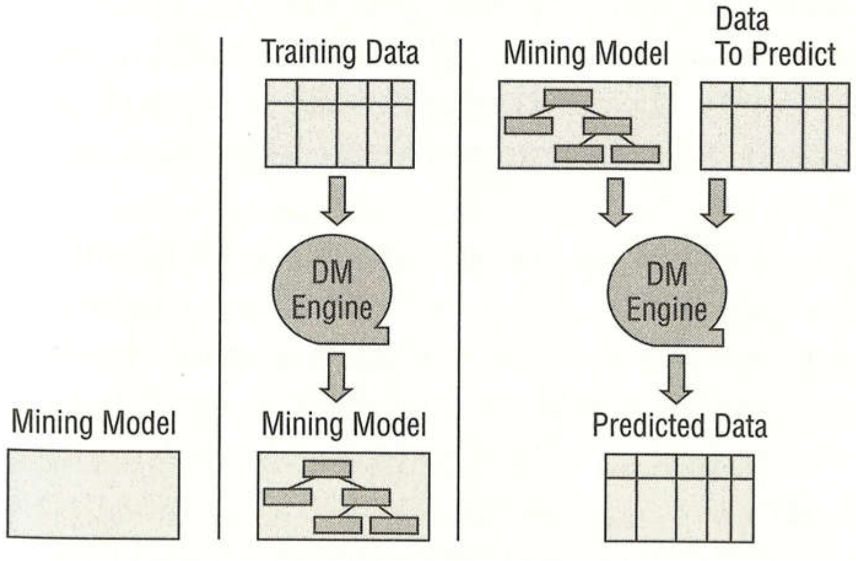
The DM Process – MS's view (DMX)

The Basics

- You select an algorithm, show the algorithm some examples called training examples and from these examples, the algorithm can extract patterns that can then be used for inspection or to deduce information about new sample input
- So the trick is forming these training examples un such a way that they are useful, informative, and accurate

The Three-Step Dance

- Creation
- Training
- prediction



Few Important Concepts

- Case – captures the traditional view of an “observation” by machine learning algorithms as consisting of all information known about a basic entity being analyzed for mining
- Case is not the same as a row in a relational table
- Information in a case is organized as attributes
- Examples of attributes are
  - Transaction ID
  - Name
  - ***Purchases*** (This is different database column)

More on Attributes

- Attributes are not exactly the same as the ones in relations
- Categorical (discrete) attributes – with a set of values, such as gender
- When the values are integers, such a number of students in classes, we also call them discrete
- Continuous attributes – numerical, such as age, or final score
- Other derived from top three – such as discretized
- Remember, the ore attributes you have, the more examples are needed to extract information from these attributes (to learn)

One More on Attribute

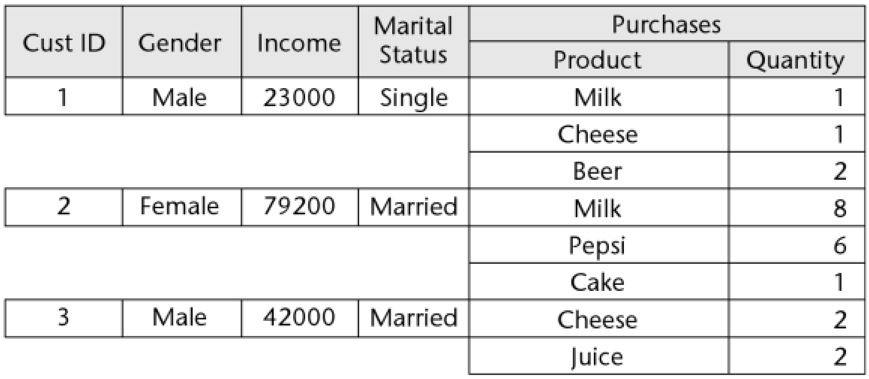
- Make attribute meaningful
    - Street address may not be meaningful, except a very few famous streets
    - Zip code can be meaningful
    - The distance from a house to a landmark is meaningful

The state and direction of an attribute

- The set of possible values of a discrete attribute is the state
    - Remember DM does not understand the inner meanings of values. For example, Divorced and Widowed both mean single
    - Limit the size of the state whenever possible
- Missing and existing are auto added states for every ttribute
- An attribute can be an input, output or both (the direction of an attribute)
- Example, of "both" is to predict if a customer would purchase a bike, even that info is provided, in "What-if" analysis
    - MS algorithms never use the output attributes to predict themselves. That is way, the predicated results and the actual *may not be* the same

Nested Case

- A case can also have table column or a nested case – example



Cust ID	Gender	Income	Marital Status	Purchases	
				Product	Quantity
1	Male	23000	Single	Milk	1
				Cheese	1
				Beer	2
2	Female	79200	Married	Milk	8
				Pepsi	6
				Cake	1
3	Male	42000	Married	Cheese	2
				Juice	2

Two Types of Keys

- Case Key – IDs each entity represented by a case
- Nested Key – ID a row in a nested ccase
- For the previous example
  - Cust ID is the Case key
  - Product may be a nested key
  - Or Product+IsOnSale may be the nested key (in the case one customer can only buy two on sale milk)

The Main Tools

- SQL Server 2014/2016
  - RDBMS to store data and support OLE DB for DM
- Visual Studio 2013
  - Provides and IDE
  - The direction MS made to use VS for all its development activity tasks, including DB, certainly include DW and DM

Data Source and Data Source View

- Data Source
  - Is the connection string indicating where the database for training and testing data is
- Data Source View (DSV)
  - An abstract layout that allows you to abstract the data in your data source

Data Source

- Two issues: file location and security
- File location

  - In the case the source data are not SQL Server DB object, you need to make sure the file path and name agree with the testing server
  - Generally, we need to move the data into SQL server
- Security
  - Most users run into problems here

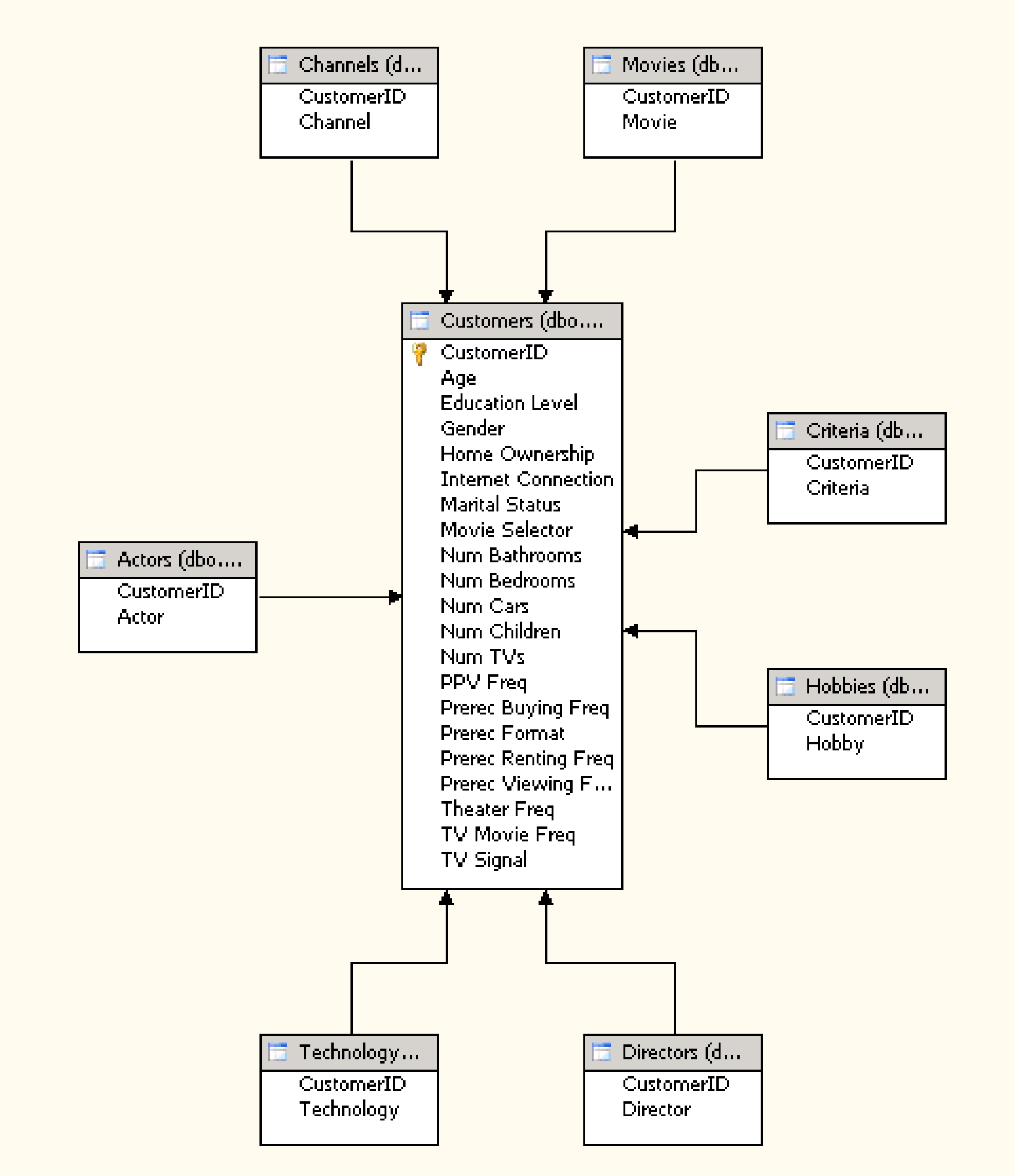
Data Source Security

- Impersonate Account
  - Make everyone's life easier – my first choice, most of the cases
- Impersonate Current User
  - Not easy to make it work
  - Causing problems when delegation is needed
- Impersonate service account
  - ***NOT*** recommended
- Default
  - Cannot control this one – therefore, not recommended

Data Source View (DSV)

- This is where the modeling begin
- Defines how you want to see the data at the data source
- Here we define case table, nested case tables, and oter lookup tables (also called dimension tables)

A different example of DSV



Named Calculation

- Create some calculated columns without affecting the source data
  - For example, fro DOB to age
  - Sharpen your SQL skills a bit because they are actually SQL calculations

Named Queries

- Again, create different views of your data without affecting the source data
- Basically SQL statements again, this time, they are complete retrieve (select) statements with join/distinct/group by etc.

Mining Structure

- The mining structure is a data structure that defines the data domain from which mining models are built
- A single mining structure can contain multiple mining exercises (mining models)
- The building blocks of mining structure are the mining structure columns, which describe the data that the data source contains
- These columns contain information such as *data style, content type* and how the data is distributed
- A mining structure can also contain nested tables. A nested table represents a one-to-many relationship between the entity of a case and its related attributes
- The mining structure does not contain information about how columns are used for a specific mining model, or about the type of algorithm that is used to build the model

Mining structure Code Example

```
CREATE MINING STRUCTURE [People3]
(
        [CustID]LONG    KEY,
        [Name] TEXT     DISCRETE,
        [Gender]        TEXT    DISCRETE,
        [Age]           LONG    CONTINUOUS,
        [AgeDisc]       LONG    DISCRETIZED(EQUAL_AREAS,3),
        [CarMake]       TEXT    DISCRETE,
        [CarModel]      TEXT    DISCRETE,
        [Purchases]     TABLE
        (
                [Product]       TEXT    KEY,
                [Quantity]      LONG    CONTINUOUS,
                [OnSale]        BOOLEAN         DISCRETE
        ),
        [Movie Ratings] TABLE
        (
                [Movie]TEXT     KEY,
                [Rating]LONG    CONTINUOUS
        )
) WITH HOLDOUT(30 PERCENT OR 10000 CASES)
```

Mining Structure Content Types

DISCRETE

The column contains discrete values.

CONTINUOUS

Calls for a continuous set of numeric data, such as income. Can be infinite in possible values: Date, Double, and Long.

DISCRETIZED

The column contains values that represent groups, or buckets, of values that are derived from a continuous column. The buckets are treated as ordered and discrete values. Data types: Date, Double, Long, and Text.

KEY

The column uniquely identifies a row. Data types: Date, Double, Long, and Text.

KEY SEQUENCE

The column is a specific type of key where the values represent a sequence of events. The values are ordered and do not have to be an equal distance apart. Data types: Date, Double, Long, and Text.

KEY TIME

The column is a specific type of key where the values represent values that are ordered and that occur on a time scale. Data types: Double, Long, and Date.

ORDERED

The column contains values that define an ordered set. However, the ordered set does not imply any distance or magnitude relationship between values in the set. For example, if an ordered attribute column contains information about skill levels in rank order from one to five, there is no implied information in the distance between skill levels; a skill level of five is not necessarily five times better than a skill level of one. Considered to be discrete in terms of content type.

CYCLICAL

The column contains values that represent a cyclical ordered set. For example, the numbered days of the week is a cyclical ordered set, because day number one follows day number seven.

- Cyclical columns are considered both ordered and discrete in terms of content type.

Continuous vs. Discrete Variables

- Continuous variable can take on any value within the resolution
    - Temperature
    - Distance
    - Weight
- Discrete variables have fixed values (even numerical)
    - Number of people
    - Number of wheels on a vehicle

Data Mining Model

- Like the mining structure, the mining model contains columns.
- A mining model is contained within the mining structure, and inherits all the values of the properties that are defined by the mining structure.
- The model can use all the columns that the mining structure contains or a subset of the columns.
- In addition to the parameters that are defined on the mining structure, the mining model contains two properties: Algorithm and Usage. The algorithm parameter is defined on the mining model, and the usage parameter is defined on the mining model column.
    - ***algorithm***
        - A model property that defines the algorithm that is used to create the model.
    - ***usage***
- A model column property that defines how a column is used by the model. You can define columns to be input columns, key columns, or predictable columns.
- A mining model is just an empty object until it is processed. When you process a model, the data that is defined by the structure is passed through the algorithm. The algorithm identifies rules and patterns within the data, and then uses these rules and patterns to populate the model.
- After you have processed a model, you can explore it by using the custom viewers that are provided in SQL Server Data Tool and SQL Server Management Studio, or by querying the model to perform predictions.
- You can create multiple models on the same structure.
- All models built from the same structure must be from the same data source. However, the models can differ as to which columns are used, how the columns are used, the type of algorithm that is used to create each model, and the parameter settings for each algorithm.

Coding Sample of Model

```
ALTER MINING STRUCTURE [People3]
ADD MINING MODEL [PredictGenderNested-Trees]
(
        [CustID],
        [Gender] PREDICT,
        [Age],
        [Purchases]
        (
                [Product],
                [Quantity],
                [OnSale]
        ),
        [Movie Ratings]
        (
                [Movie],
                [Rating]
        )
) USING Microsoft_Decision_Trees(COMPLEXITY_PENALTY=0.5)
WITH FILTER(Age > 20)
```

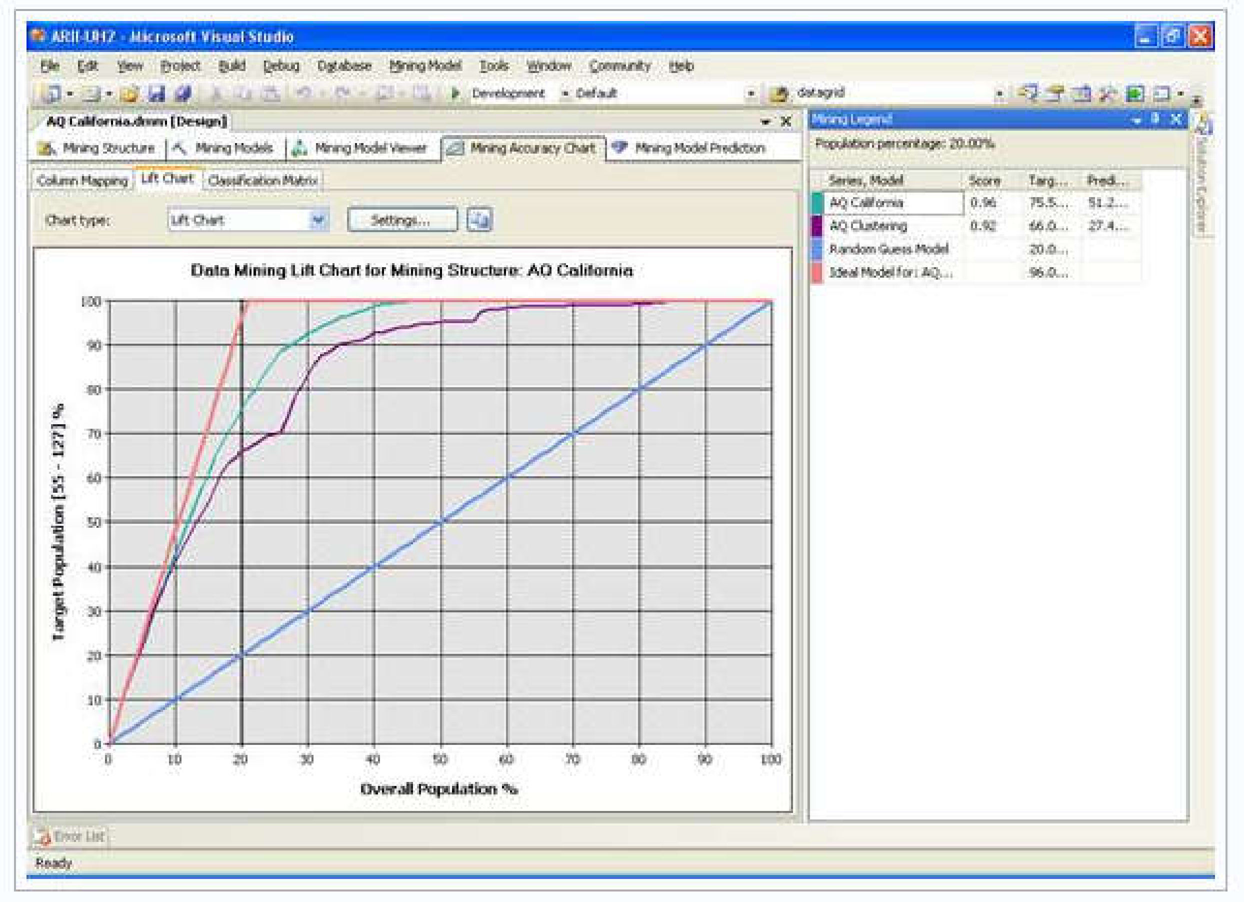
Predict

- PREDICT
  - This column can be predicted by the model, and it can be supplied in input cases to predict the value of other predictable columns.
- PREDICT_ONLY
  - This column can be predicted by the model, but its values cannot be used in input cases to predict the value of other predictable columns.
- You can have multiple columns marked as predict

Accuracy Chart

- Lift Chart
  - One line for each model
  - A random line
  - Ideal line
- Lift is a measure of the effectiveness of a predictive model
- You can also perform profit calculation

Example of a Lift Chart



Population percentage: 20.00%

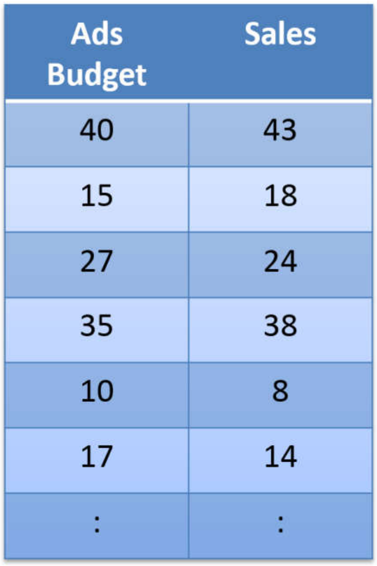
	Series, Model	Score	Targ...	Predi...
🟩	AQ California	0.96	75.5...	51.2...
🟪	AQ Clustering	0.92	66.0...	27.4...
🟦	Random Guess Model		20.0...	
🟥	Ideal Model for: AQ...		96.0...	

Executing Queries against the Model

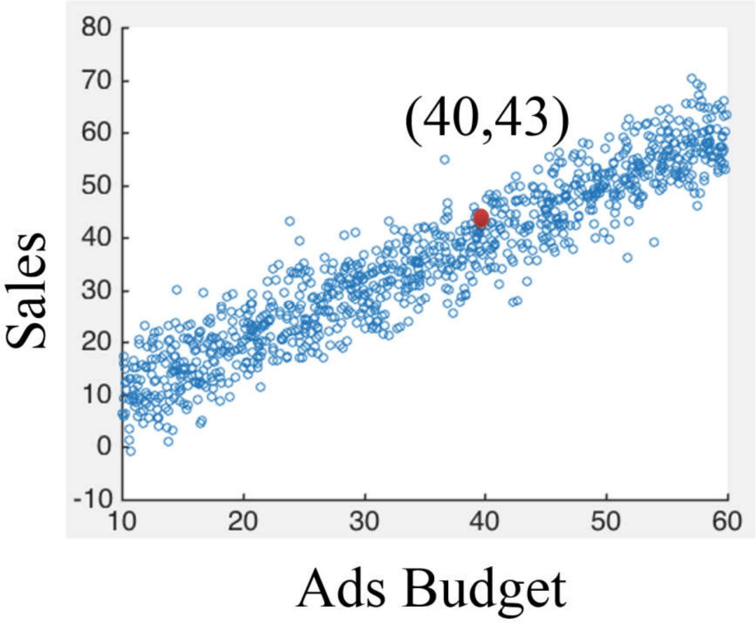
- Using the Mining Model prediction tab
- Can modify the SQL statements directly
- You can save the result as a database table

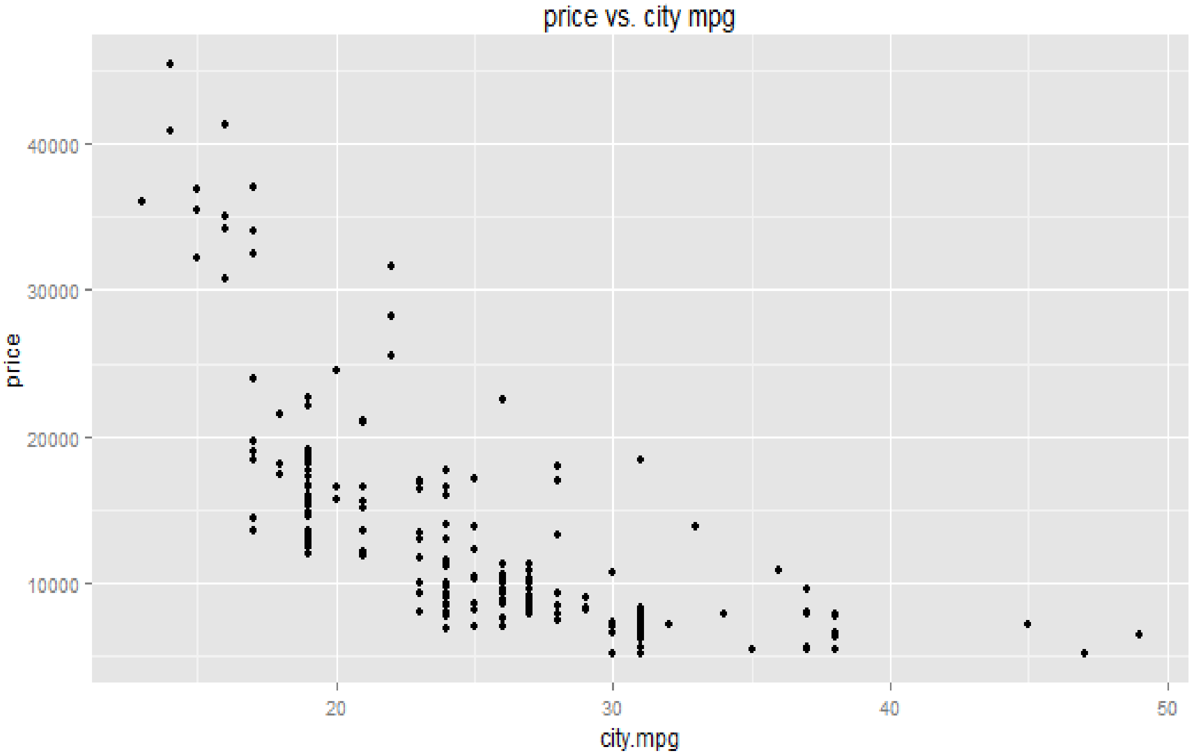
Types of Plots

Scatter Plots

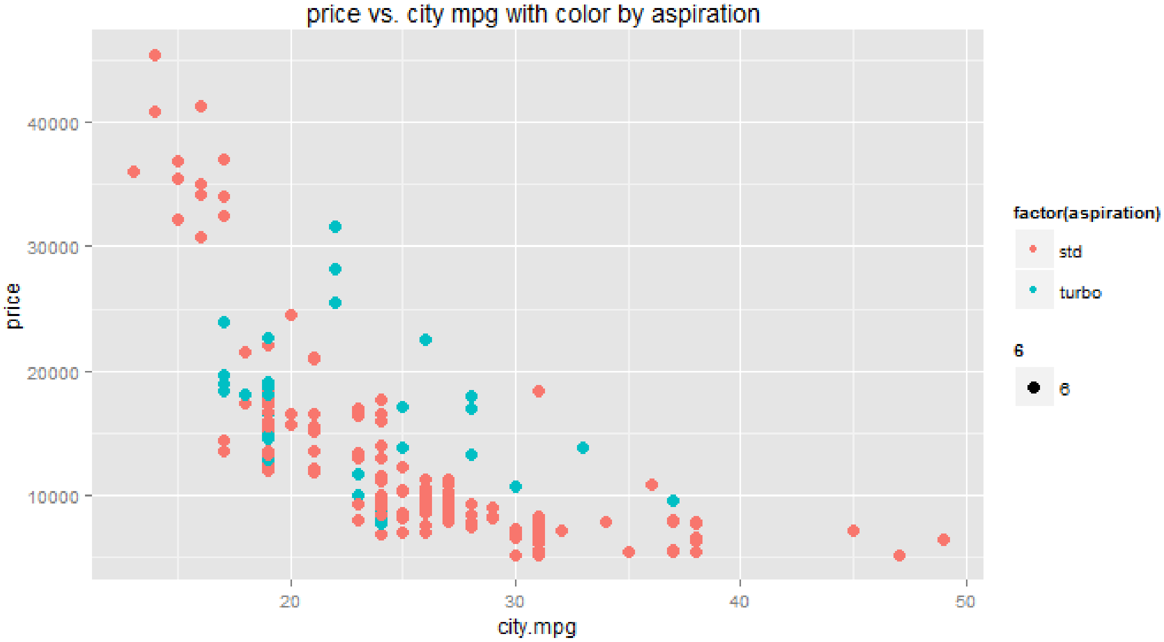


Ads Budget	Sales
40	43
15	18
27	24
35	38
10	8
17	14
:	:

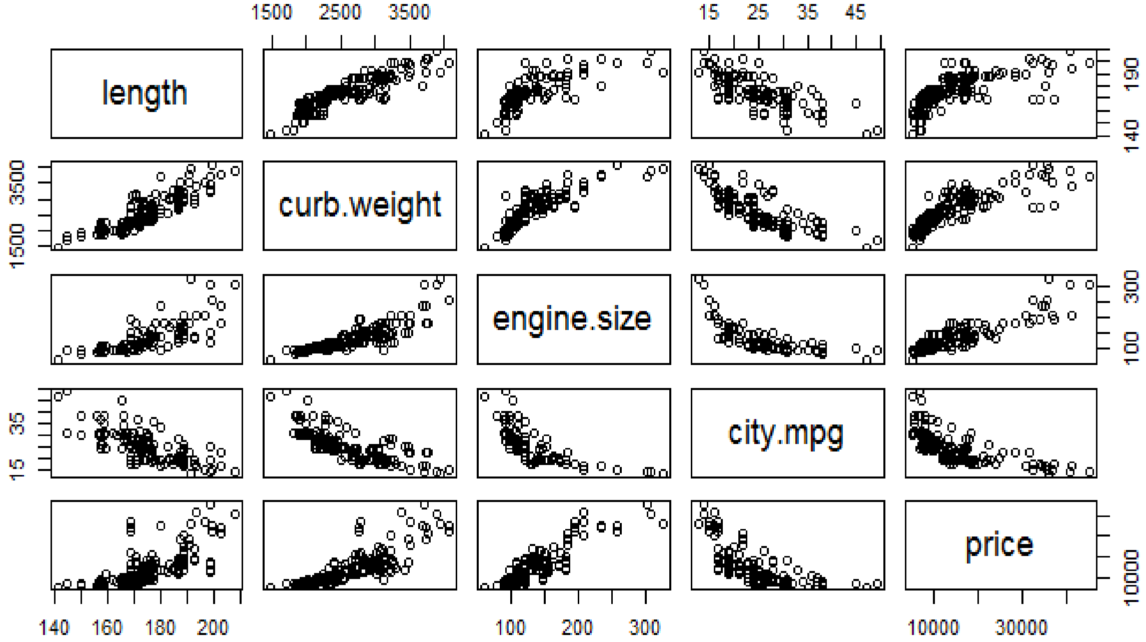




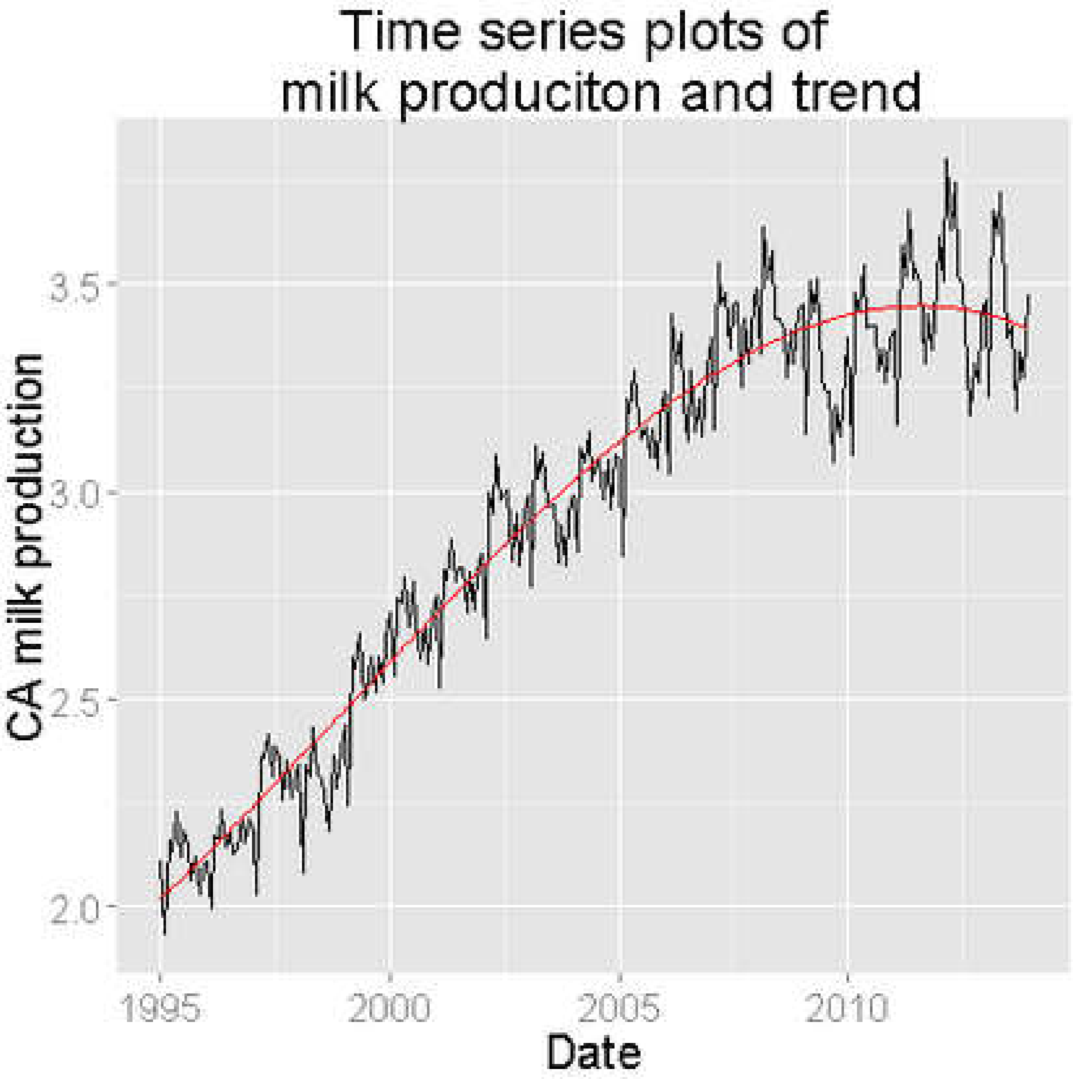
Scatter Plot (+Color by Category)



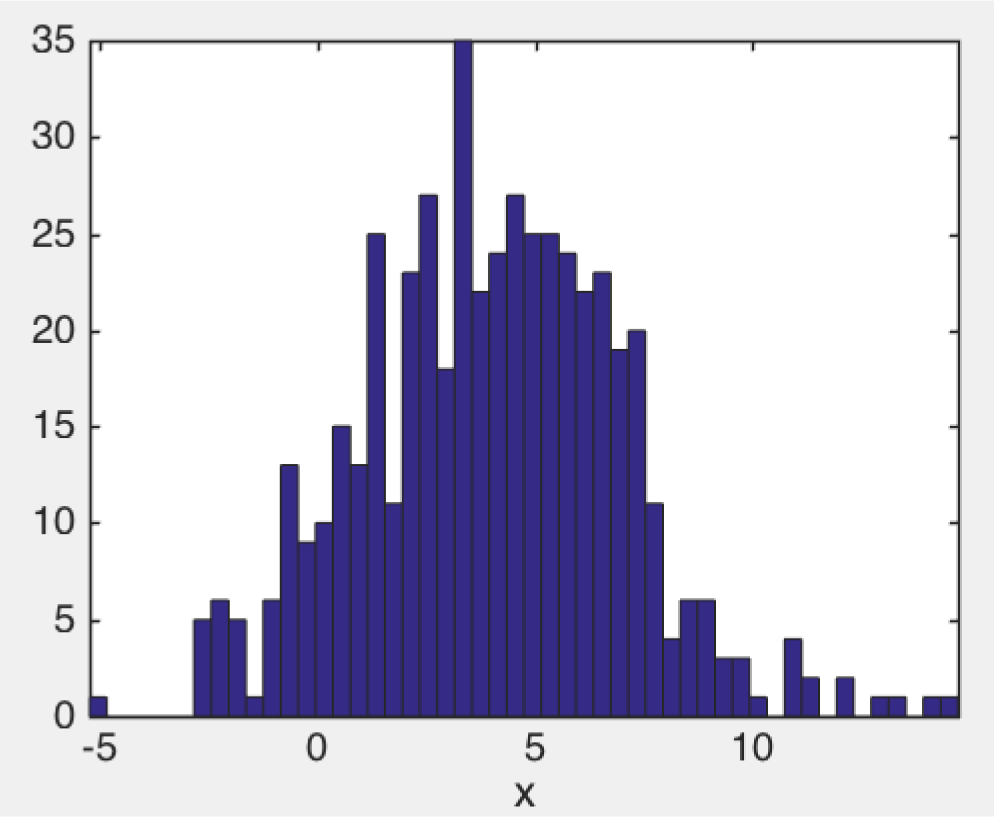
Scatter Ploy Matrix



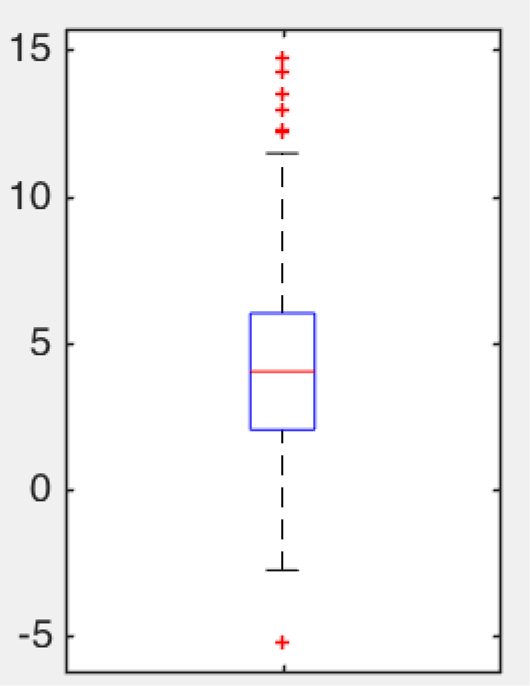
Line Plot


Time series plots of
milk produciton and trend
CA milk production
3.5
3.0
2.5
2.0
1995
2000
2005
2010
Date

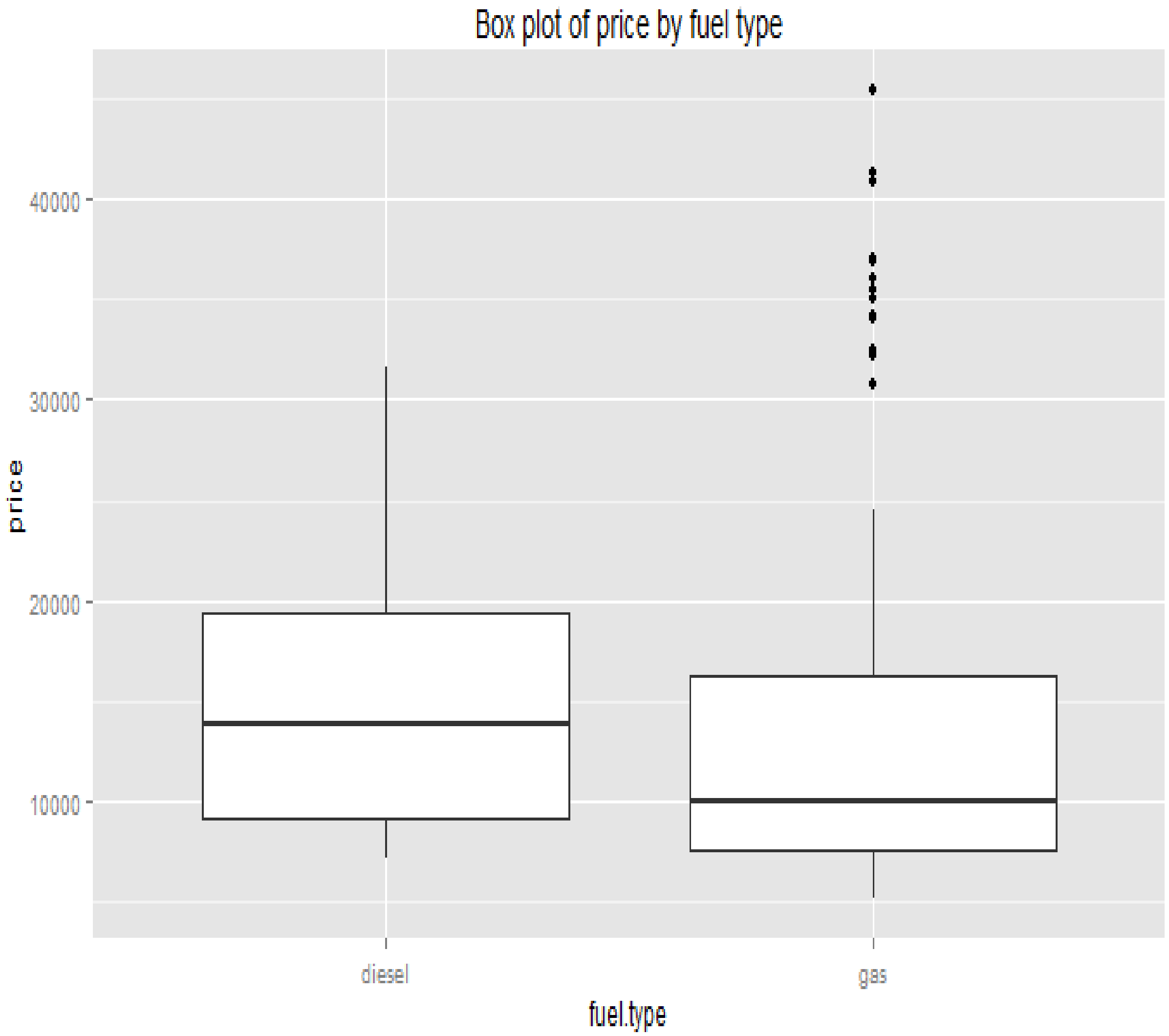
Histogram



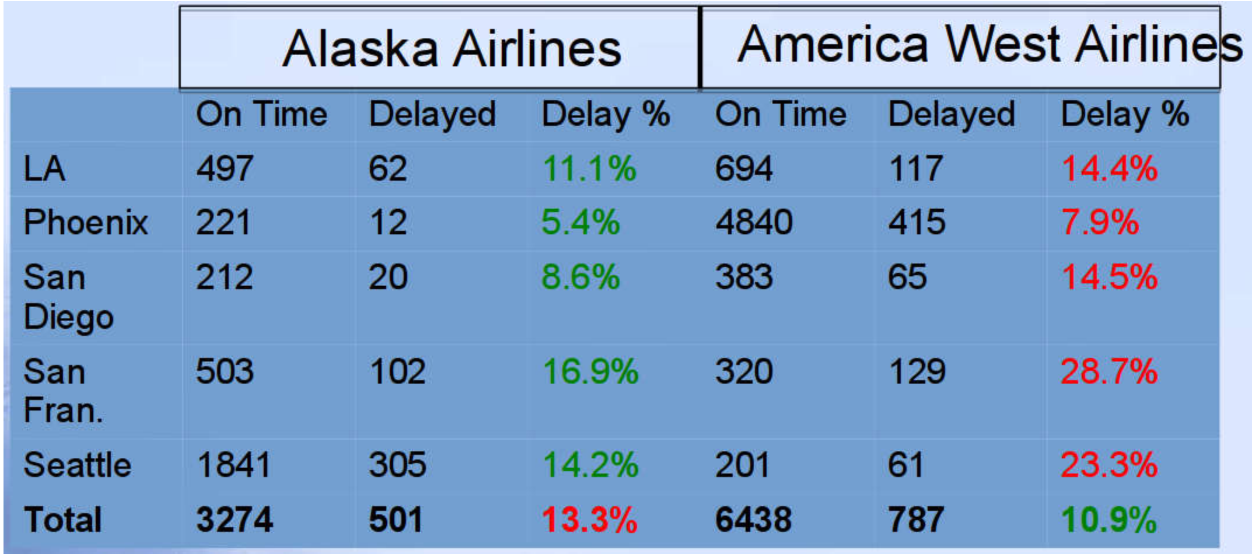
Box Plot



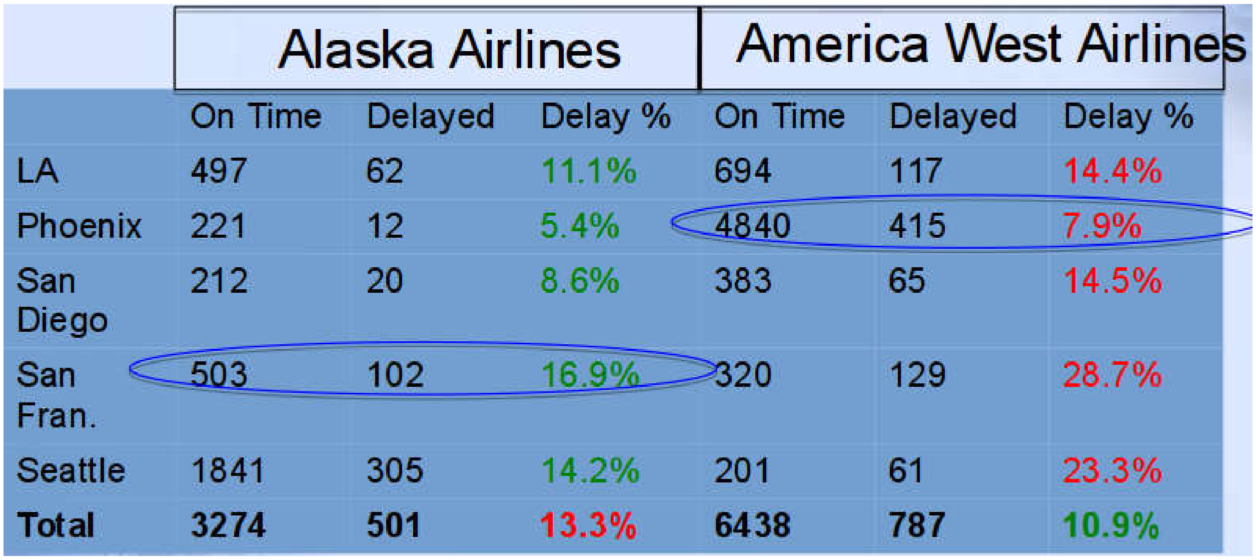
Box Plot (group by category)



Simpson’s Paradox



	Alaska Airlines			America West Airlines		
	On Time	Delayed	Delay %	On Time	Delayed	Delay %
LA	497	62	11.1%	694	117	14.4%
Phoenix	221	12	5.4%	4840	415	7.9%
San Diego	212	20	8.6%	383	65	14.5%
San Fran.	503	102	16.9%	320	129	28.7%
Seattle	1841	305	14.2%	201	61	23.3%
Total	**3274**	**501**	**13.3%**	**6438**	**787**	**10.9%**



	Alaska Airlines			America West Airlines		
	On Time	Delayed	Delay %	On Time	Delayed	Delay %
LA	497	62	11.1%	694	117	14.4%
Phoenix	221	12	5.4%	4840	415	7.9%
San Diego	212	20	8.6%	383	65	14.5%
San Fran.	503	102	16.9%	320	129	28.7%
Seattle	1841	305	14.2%	201	61	23.3%
Total	**3274**	**501**	**13.3%**	**6438**	**787**	**10.9%**