CS(STAT)5525 : Data Analytics
Lecture #2

Reza Jafari, Ph.D

Collegiate Associate Professor
rjafari@vt.edu

COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH.

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics.

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics.
- Key motivations of data exploration include.

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics.
- Key motivations of data exploration include.
  - Helping to select the right tool for preprocessing or analysis.

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics.
- Key motivations of data exploration include.
  - Helping to select the right tool for preprocessing or analysis.
  - Making use of humans' abilities to recognize patterns.

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics.
- Key motivations of data exploration include.
  - Helping to select the right tool for preprocessing or analysis.
  - Making use of humans' abilities to recognize patterns.
  - People can recognize patterns not captured by data analysis tools.

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics.
- Key motivations of data exploration include.
  - Helping to select the right tool for preprocessing or analysis.
  - Making use of humans' abilities to recognize patterns.
  - People can recognize patterns not captured by data analysis tools.
- Related to the area of Exploratory Data Analysis (EDA)

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics.
- Key motivations of data exploration include.
  - Helping to select the right tool for preprocessing or analysis.
  - Making use of humans' abilities to recognize patterns.
  - People can recognize patterns not captured by data analysis tools.
- Related to the area of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics.
- Key motivations of data exploration include.
  - Helping to select the right tool for preprocessing or analysis.
  - Making use of humans' abilities to recognize patterns.
  - People can recognize patterns not captured by data analysis tools.
- Related to the area of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey
  - Seminal book is Exploratory Data Analysis by Tukey

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics.
- Key motivations of data exploration include.
  - Helping to select the right tool for preprocessing or analysis.
  - Making use of humans' abilities to recognize patterns.
  - People can recognize patterns not captured by data analysis tools.
- Related to the area of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey
  - Seminal book is Exploratory Data Analysis by Tukey
  - A nice online introduction can be found in Chapter 1 of the NIST/SEMATECH e-Handbook of Statistical Methods [Web Link]

Data Analysis Approaches

- The three popular data analysis approaches are :

1- Classical Analysis

The data collection is followed by a model (normality, linearity,etc) and then analysis, estimation and testing that follows are focused on the parameters of that model.

2- EDA

For EDA, the data is followed immediately by analysis with a goal of inferring what model would be appropriate.

Data Analysis Approaches

- The <u>three</u> popular data analysis approaches are :
  1. Classical Analysis

1- Classical Analysis

The data collection is followed by a model (normality, linearity,etc) and then analysis, estimation and testing that follows are focused on the parameters of that model.

- Problem → Data → Model → Analysis → Conclusions

2- EDA

For EDA, the data is followed immediately by analysis with a goal of inferring what model would be appropriate.

Data Analysis Approaches

- The three popular data analysis approaches are :
  1. Classical Analysis
  2. EDA

1- Classical Analysis

The data collection is followed by a model (normality, linearity,etc) and then analysis, estimation and testing that follows are focused on the parameters of that model.

- Problem $\rightarrow$ Data $\rightarrow$ Model $\rightarrow$ Analysis $\rightarrow$ Conclusions

2- EDA

For EDA, the data is followed immediately by analysis with a goal of inferring what model would be appropriate.

- Problem $\rightarrow$ Data $\rightarrow$ Analysis $\rightarrow$ Model $\rightarrow$ Conclusions

Data Analysis Approaches

- The three popular data analysis approaches are :
  1. Classical Analysis
  2. EDA
  3. Bayesian

1- Classical Analysis

The data collection is followed by a model (normality, linearity,etc) and then analysis, estimation and testing that follows are focused on the parameters of that model.

- Problem $\rightarrow$ Data $\rightarrow$ Model $\rightarrow$ Analysis $\rightarrow$ Conclusions

2- EDA

For EDA, the data is followed immediately by analysis with a goal of inferring what model would be appropriate.

- Problem $\rightarrow$ Data $\rightarrow$ Analysis $\rightarrow$ Model $\rightarrow$ Conclusions

Data Analysis Approaches

3- Bayesian

For Bayesian analysis, the analyst attempts to in corporate scientific/engineering knowledge/experitise into the analysis by imposing a data independent distribution on the parameters of the selected model.

- Problem $\rightarrow$ Data $\rightarrow$ Model $\rightarrow$ Prior Distribution $\rightarrow$ Analysis $\rightarrow$ Conclusions

Data Analysis Approaches

3- Bayesian

For Bayesian analysis, the analyst attempts to in corporate scientific/engineering knowledge/experitise into the analysis by imposing a data independent distribution on the parameters of the selected model.

- Problem → Data → Model → Prior Distribution → Analysis → Conclusions

- In the real world, data analysts freely mix elements of all the above three approaches (and other approaches).

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization.

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization.
  - Clustering & anomaly detection were viewed as exploratory techniques.

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization.
  - Clustering & anomaly detection were viewed as exploratory techniques.
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory.

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization.
  - Clustering & anomaly detection were viewed as exploratory techniques.
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory.
- In our discussion of data exploration, we focus on:

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization.
  - Clustering & anomaly detection were viewed as exploratory techniques.
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory.
- In our discussion of data exploration, we focus on:
  - Summary statistics

Techniques Used In Data Exploration

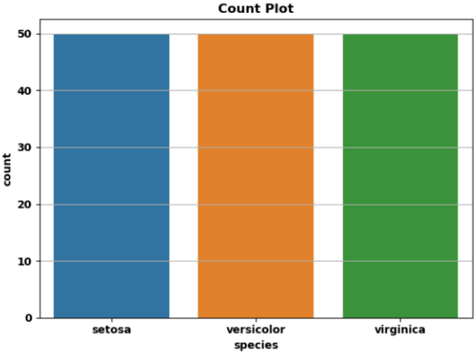
- In EDA, as originally defined by Tukey
  - The focus was on visualization.
  - Clustering & anomaly detection were viewed as exploratory techniques.
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory.
- In our discussion of data exploration, we focus on:
  - Summary statistics
  - Visualization

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization.
  - Clustering & anomaly detection were viewed as exploratory techniques.
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory.
- In our discussion of data exploration, we focus on:
  - Summary statistics
  - Visualization
  - Online Analytical Processing (OLAP)

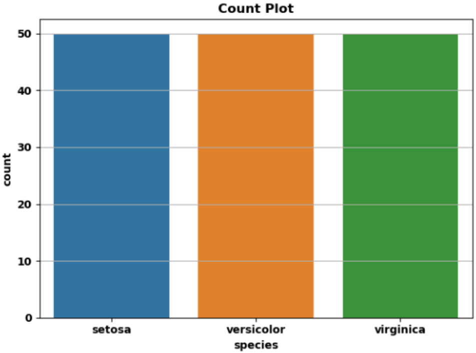
Iris Sample Data set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.



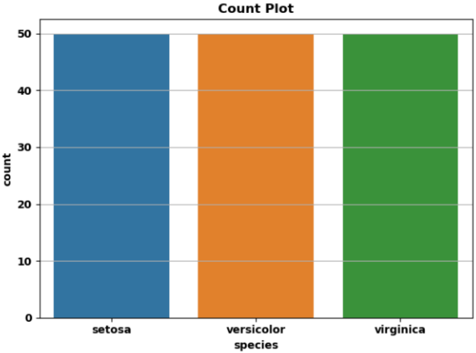
Iris Sample Data set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
- The iris dataset can be obtained from the seaborn package in python using sns.load_dataset('iris')



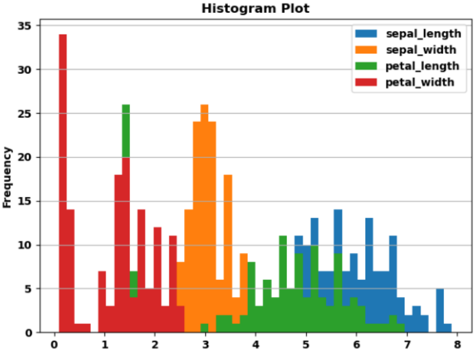
Iris Sample Data set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
- The iris dataset can be obtained from the seaborn package in python using sns.load_dataset('iris')
- Four attributes sepal width & length, petal width & length



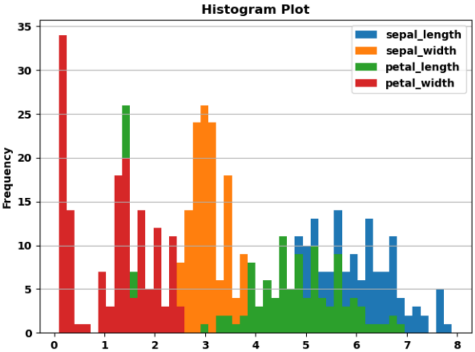
Visualization techniques: Histogram

- Histogram



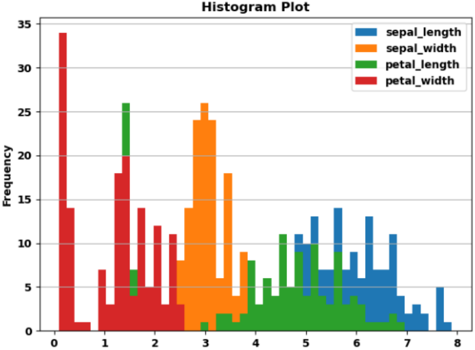
Visualization techniques: Histogram

- Histogram
  - Usually shows the distribution of values of a single variable



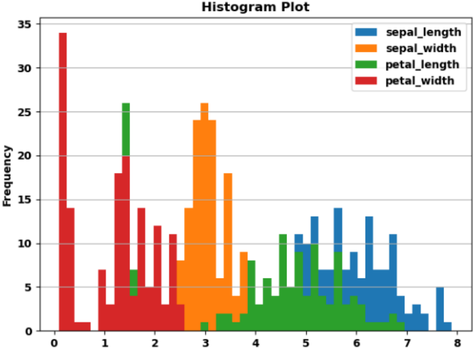
Visualization techniques: Histogram

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.



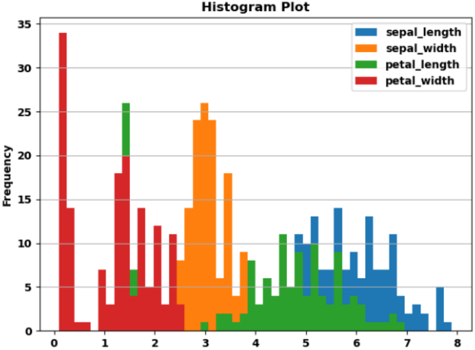
Visualization techniques: Histogram

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the frequency of objects.



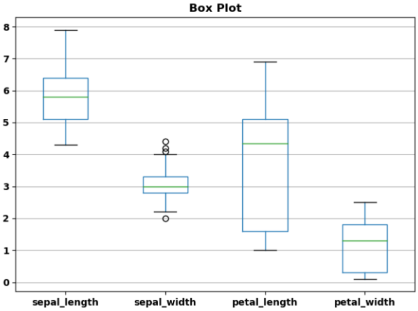
Visualization techniques: Histogram

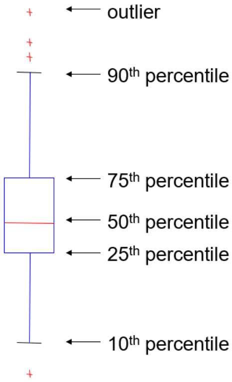
- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the frequency of objects.
  - Shape of histogram depends on the number of bins



Visualization techniques: Histogram

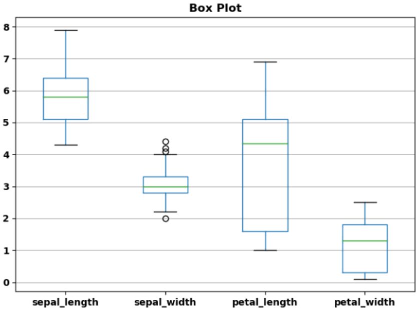
- Boxplot invented by J. Tukey

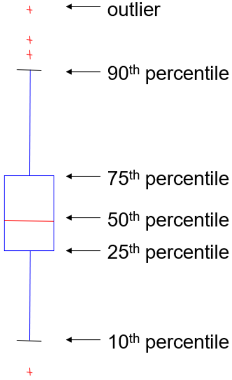




Visualization techniques: Histogram

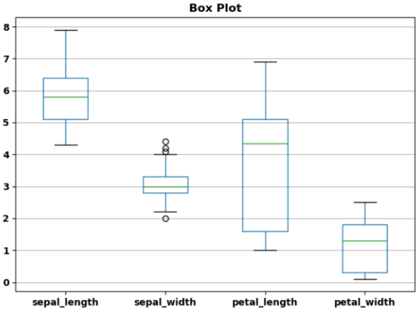
- Boxplot invented by J. Tukey
  - Usually shows the distribution of values of a single variable

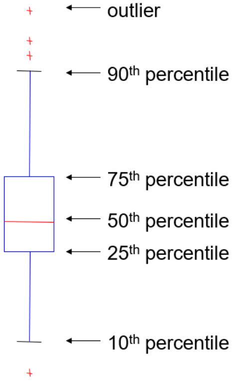




Visualization techniques: Histogram

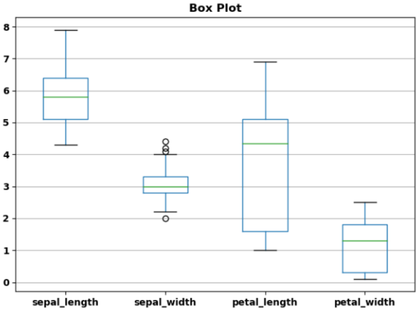
- Boxplot invented by J. Tukey
  - Usually shows the distribution of values of a single variable
  - Another way of displaying the distribution of data

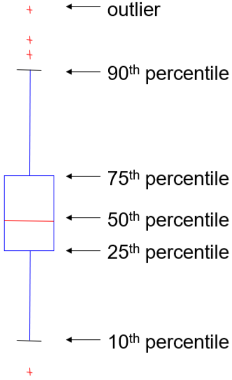




Visualization techniques: Histogram

- Boxplot invented by J. Tukey
  - Usually shows the distribution of values of a single variable
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot





Visualization techniques: Scatter Plots

- Scatter plots

Visualization techniques: Scatter Plots

- Scatter plots
  - Attributes values determine the position

Visualization techniques: Scatter Plots

- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots

Visualization techniques: Scatter Plots

- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects

Visualization techniques: Scatter Plots

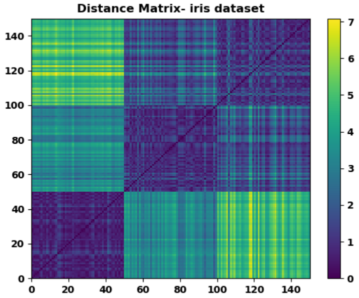
- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes

Visualization techniques: Scatter Plots

- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
  - See example on the next slide

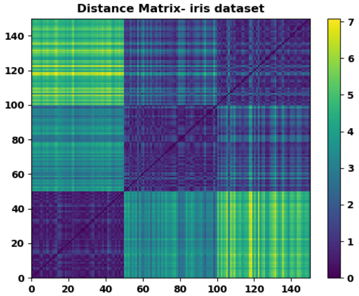
Visualization techniques: Distance Matrix

- We can see 3 noticeable clusters:



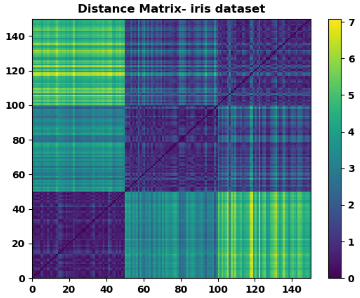
Visualization techniques: Distance Matrix

- We can see 3 noticeable clusters:
  1. one which is rather dense (bottom left) and far from the others.



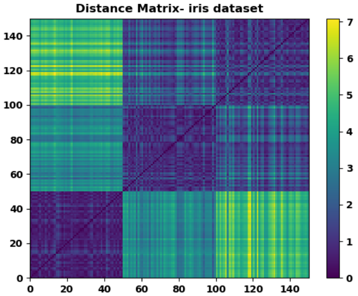
Visualization techniques: Distance Matrix

- We can see 3 noticeable clusters:
  1. one which is rather dense (bottom left) and far from the others.
  2. two which are quite close but differ in their respective distance to the third one (bottom left).



Visualization techniques: Distance Matrix

- We can see 3 noticeable clusters:
  1. one which is rather dense (bottom left) and far from the others.
  2. two which are quite close but differ in their respective distance to the third one (bottom left).
- These three clusters refers to 3-types of flowers ('setosa', 'versicolor', 'virginica')



Visualization techniques: Distance Matrix-python practice

- scipy.spatial.distance.pdist() returns the pairwise distances between observations in n-dimensional space.

Visualization techniques: Distance Matrix-python practice

- scipy.spatial.distance.pdist() returns the pairwise distances between observations in n-dimensional space.
- A condensed distance matrix as returned by pdist can be converted to a full distance matrix by using scipy.spatial.distance.squareform().

Visualization techniques: Distance Matrix-python practice

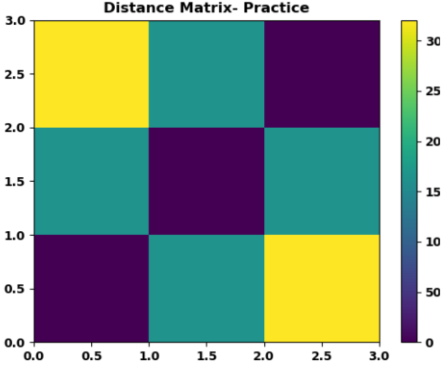
- scipy.spatial.distance.pdist() returns the pairwise distances between observations in n-dimensional space.
- A condensed distance matrix as returned by pdist can be converted to a full distance matrix by using scipy.spatial.distance.squareform().
- Lets create a normally distributed systematic dataset with 1000 observations that represent 3 clusters.

Visualization techniques: Distance Matrix-python practice

- scipy.spatial.distance.pdist() returns the pairwise distances between observations in n-dimensional space.
- A condensed distance matrix as returned by pdist can be converted to a full distance matrix by using scipy.spatial.distance.squareform().
- Lets create a normally distributed systematic dataset with 1000 observations that represent 3 clusters.
  - Cluster 1: mean = 0 variance 1
  - Cluster 1: mean = 5 variance 1
  - Cluster 1: mean = 10 variance 1

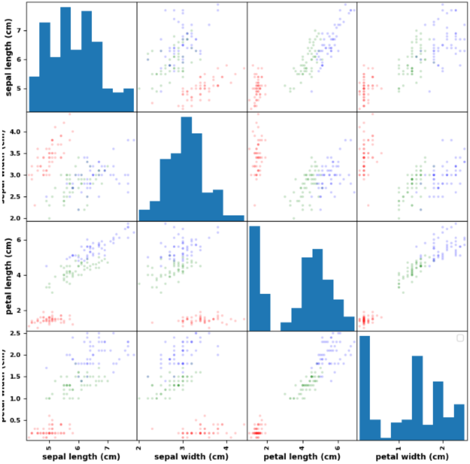
Visualization techniques: Distance Matrix-python practice

```
import numpy as np
from scipy.spatial.distance import pdist, squareform
import matplotlib.pyplot as plt
np.random.seed(123)

x1 = np.random.normal(0,1,1000)
x2 = np.random.normal(5,1,1000)
x3 = np.random.normal(10,1,1000)

X = np.vstack((x1,x2,x3))
dist_mat = squareform(pdist(X))

N = len(X)
plt.pcolormesh(dist_mat)
plt.colorbar()
plt.xlim([0,N])
plt.ylim([0,N])
plt.title('Distance Matrix- Practice')
plt.show()
```



Visualization techniques: Scatter Plots

```
from sklearn.datasets import load_iris
from numpy import array
from pandas import DataFrame
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt

iris = load_iris()
df = DataFrame(iris.data, columns=iris.feature_names)
colors=array(50*['r']+50*['g']+50*['b'])
scatter_matrix(df, alpha=0.2, figsize=(10,10),
               color=colors)
plt.legend()
plt.show()
```



Online Analytical Processing

- OLAP (Online Analytical Processing) is the technology behind Business Intelligence (BI) applications.

Online Analytical Processing

- OLAP (Online Analytical Processing) is the technology behind Business Intelligence (BI) applications.
- OLAP is a powerful technology for data discovery, including capabilities for limitless report, viewing, complex analytical calculations and predictive "what if scenario" (budget, forecast) planing.

Online Analytical Processing

- OLAP (Online Analytical Processing) is the technology behind Business Intelligence (BI) applications.
- OLAP is a powerful technology for data discovery, including capabilities for limitless report, viewing, complex analytical calculations and predictive "what if scenario" (budget, forecast) planing.
- OLAP technology has been defined as the ability to achieve fast access to shared multidimensional information.

Online Analytical Processing

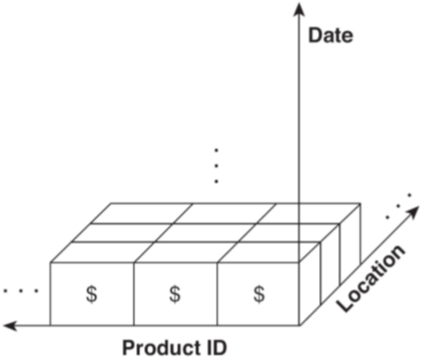
- OLAP (Online Analytical Processing) is the technology behind Business Intelligence (BI) applications.
- OLAP is a powerful technology for data discovery, including capabilities for limitless report, viewing, complex analytical calculations and predictive "what if scenario" (budget, forecast) planing.
- OLAP technology has been defined as the ability to achieve fast access to shared multidimensional information.
- Unlike relational databases, two-dimensional row-by-column format, OLAP used **Cubes** terminology to store arrays of consolidated information.

Online Analytical Processing

- OLAP (Online Analytical Processing) is the technology behind Business Intelligence (BI) applications.
- OLAP is a powerful technology for data discovery, including capabilities for limitless report, viewing, complex analytical calculations and predictive "what if scenario" (budget, forecast) planing.
- OLAP technology has been defined as the ability to achieve fast access to shared multidimensional information.
- Unlike relational databases, two-dimensional row-by-column format, OLAP used **Cubes** terminology to store arrays of consolidated information.
- The data and formulas are stored in an optimized multidimensional database, while views of the data are created on demand.

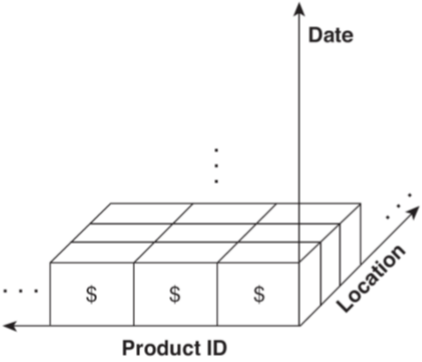
Data Cube Example

- Consider a data set that records the sales of products at a number of company stores at various dates.



Data Cube Example

- Consider a data set that records the sales of products at a number of company stores at various dates.
- This data can be represented as a 3-dimensional array.



Data Cube Example

- The following figure table shows one of the two-dimensional aggregations, along with two of the one-dimensional aggregation and the overall total.

date

product ID	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
1	$1,001	$987	...	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	...	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	...	$631,221	$227,352,127

Online Analytical Processing