# DM LAB 2

BY

ENG. JOUD KHATTAB

# 2 INTRODUCTION

- Database Review
- Case Study:
  - Project Scope
  - DB Diagram
  - Identify Possible Knowledge
  - Ex: Library Management System, Health Center System
- Business Problems for Data Mining
- Data Mining Tasks
- Data Collection & Cleaning

# 3 CASE STUDY 1 LIBRARY MANAGEMENT SYSTEM

- A Library Management System is a software built to handle the primary housekeeping functions of a library.
- Libraries rely on library management systems to manage asset collections as well as relationships with their members.
- Library management systems help libraries keep track of the books and their checkouts, as well as members' subscriptions and profiles.
- Library management systems also involve maintaining the database for entering new books and recording books that have been borrowed with their respective due dates.
- The database should also contain:
  - Customers, Subscriber, books, authors, supplier, branch, …..

# 4 CASE STUDY 2 HEALTH CENTER SYSTEM

- Doctor Assistant database is a databased that can be used in children doctor's clinic to store all the information of their patients:
  - From the personal information like (name, date of birth, gender, telephone, address and many other).
  - To the information of his birth like (birth weight, birth place, birth type, …).
  - The Medical examination for the doctor (what is the disease that the patient has and what are the medicine, analysis and the nutritional recommendations that he should take).
  - Records for the Vaccines that patient took. The Family diseases if there are any.
- The database should also contain:
  - bookmark for list of Hospitals, Doctors, Vaccines, Medicines, Nutritional Recommendations, Medical Analysis and Diseases, along with their details.

# BUSINESS PROBLEMS FOR DATA MINING

Data mining techniques can be used in virtually all business applications, answering various types of businesses questions.

# 6 BUSINESS PROBLEMS FOR DATA MINING

1. Recommendation Generation
   - Ex: what products or services should you offer to your customers?
2. Churn Analysis
   - Ex: which customers are most likely to switch to a competitor?
3. Risk Management
   - Ex: should a loan be approved for a particular customer?
4. Customer Segmentation
   - Ex: how do you think of your customers?
5. Targeted Ads
   - Ex: web retailers sites like to personalize their content for their customers.
6. Forecasting
   - Ex: how many products will you sell next week in this store?

# 7 DATA MINING TASKS

1. Classification
   - Classification is the act of assigning a category to each case.
2. Clustering
   - Clustering is also called segmentation. It is used to identify natural groupings of cases based on a set of attributes.
3. Association
   - Association is also called market basket analysis.
4. Regression
   - The regression task is similar to classification, except that instead of looking for patterns that describe a class, the goal is to find patterns to determine a numerical value.

# DATA COLLECTION & CLEANING

# DATA MINING PROJECT CYCLE
# DATA COLLECTION

- Business data is stored in many systems across an enterprise.
- For example, at Microsoft, there are hundreds of online transaction processing (OLTP) databases and more than 70 data warehouses.
- The first step is to pull the relevant data into a database or a data mart where the data analysis is applied.
  - For example, if you want to analyze your website's click stream, the first step is to download the log data from your web servers.

# DATA MINING PROJECT CYCLE
# DATA COLLECTION

- Sometimes you might be lucky and find that there is already an existing data warehouse on the subject of your analysis.
- However, in many cases, the data in the data warehouse is not rich enough and must be supplemented with additional data.
  - For example, the log data from the web servers contains only data about web behavior and little (if any) data about the customers.
- You may need to gather customer information from other company systems or purchase demographic data to build models that meet your business requirements.

# DATA MINING PROJECT CYCLE
# DATA CLEANING AND TRANSFORMATION

- Data cleaning and transformation are the most resource-consuming steps in a data mining project.
- The purpose of data cleaning is to remove noise and irrelevant information from the data set.
- The purpose of data transformation is to modify the source data in ways that make it useful for mining.

# DATA MINING PROJECT CYCLE
# DATA CLEANING AND TRANSFORMATION

- Various techniques are applied to clean and transform data, including:
  1. Numerical Transformation
  2. Grouping
  3. Aggregation
  4. Missing Value Handling
  5. Removing Outliers

# DATA MINING PROJECT CYCLE
# DATA CLEANING AND TRANSFORMATION

- Numerical Transformation:
  - For continuous data such as income and age, a typical transformation is to bin (or discretize) the data into buckets.
  - For example, you may want to bin Age into five predefined age groups.
  - Additionally, continuous data is often normalized.
  - Normalization maps all numerical values to a range (such as between 0 and 1).

# DATA MINING PROJECT CYCLE
# DATA CLEANING AND TRANSFORMATION

- Grouping:
  - Discrete data often has more distinct values than are useful. You can group these values to reduce the model complexity.
  - For example, the column Profession may have many different types of engineers, such as Software Engineer, Telecom Engineer, Mechanical Engineer, and so on.
  - You can group all of these professions to the single value Engineer.

# DATA MINING PROJECT CYCLE
# DATA CLEANING AND TRANSFORMATION

- Aggregation:
  - Aggregation is an important transformation to derive additional value from your data.
  - Suppose you want to group customers based on their phone usage.
  - If the call detail record information is too detailed for the model, you must aggregate all the calls into a few derived attributes such as total number of calls and the average call duration.
  - These derived attributes can later be used in the model.

# DATA MINING PROJECT CYCLE
# DATA CLEANING AND TRANSFORMATION

- Missing Value Handling:
  - Most data sets contain missing values.
  - This can be caused by many different things.
  - For example, you may have two customer tables coming from two OLTP databases that, when merged, have missing values because the tables are not aligned.
  - Another example occurs when customers don't supply data values such as age.
  - Another is when you have stock market values with blanks because the markets are closed on weekends and holidays.

# DATA MINING PROJECT CYCLE
# DATA CLEANING AND TRANSFORMATION

- Missing Value Handling:
  - Addressing missing values is important, because it is reflected in the business value of your solution.
  - You may need to retain the missing data (for example, customers who refuse to report their age may have other interesting things in common).
  - You may need to discard the entire record (having too many unknowns could pollute your model). Or, you may simply be able to replace missing values with some other value (such as the previous value for time-series data such as stock market values, or the most popular value).

# DATA MINING PROJECT CYCLE
# DATA CLEANING AND TRANSFORMATION

- Missing Value Handling:

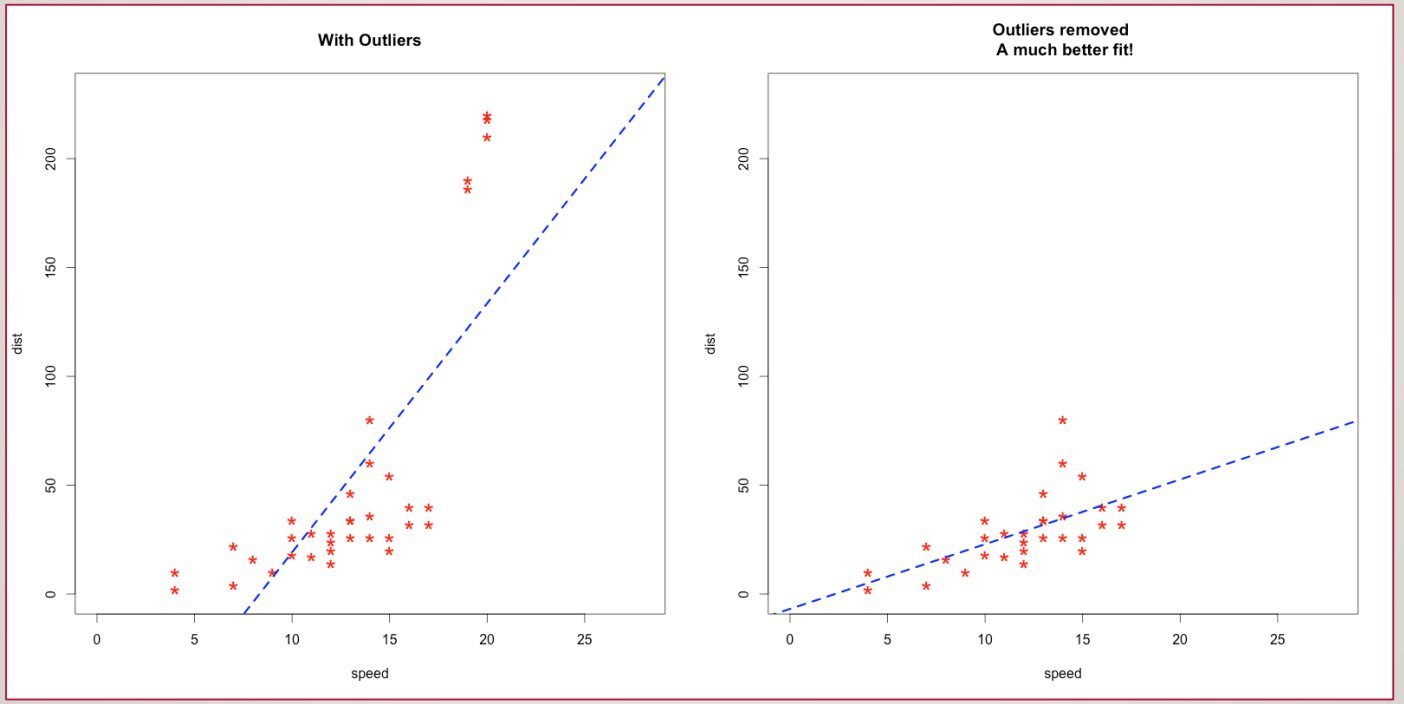
| Column 0 | age | years_seniority | income | parking_space | attending_party | entree | pets | emergency_contact |
|---|---|---|---|---|---|---|---|---|
| Tony | 48 | 27 | | 1 | 5 | shrimp | | Pepper |
| Donald | 67 | 25 | 86 | 10 | 2 | beef | | Jane |
| Henry | 69 | 21 | 95 | 6 | 1 | chicken | 62 | Janet |
| Janet | 62 | 21 | 110 | 3 | 1 | beef | | Henry |
| Nick | | 17 | | 4 | | | | |
| Bruce | 37 | 14 | 63 | | 1 | veggie | | NA |
| Steve | 83 | | 77 | 7 | 1 | chicken | | n/a |
| Clint | 27 | 9 | 118 | 9 | | shrimp | 3 | None |
| Wanda | 19 | 7 | 52 | 2 | 2 | shrimp | | empty |
| Natasha | 26 | 4 | 162 | 5 | 3 | | | _ |
| Carol | | 3 | 127 | 11 | 1 | veggie | 1 | """" |
| Mandy | 44 | 2 | 68 | 8 | 1 | chicken | | null |

# DATA MINING PROJECT CYCLE
# DATA CLEANING AND TRANSFORMATION

- Removing Outliers:
  - Outliers are abnormal data and can be real or (as is often the case) errors.
  - Abnormal data has an effect on the quality of your results.
  - The best way to deal with outliers typically is to simply remove them before beginning the analysis.
  - For example, you could remove 0.5 percent of the customers with highest or lowest income to eliminate any situations of people having negative or extremely unlikely incomes.

# DATA MINING PROJECT CYCLE
# DATA CLEANING AND TRANSFORMATION

- Removing Outliers:




# THE END