# **Final Exam Responses**

**Introduction to Data Analytics**

**IS 534**

**Instructor: Dr. Lynn Collen**



Submitted by,

Venkata Ayyappa D

ID: 13228200

## .0 Reasons why people are needed to direct Datamining?

With the recent technological advancement, most of the organizations wants to assign a history tag of its customers (preferences, behavior). In fact, huge amounts of data being collected by most of the organizational servers. What is being learned from all this data?

Data mining can be described as a process of finding anomalies, patters and correlations in large data sets to predict outcomes (Future trends). It all depends on the cooperation between the Humans and the Data mining systems. Automation of these Data Mining systems cannot be considered as a substitute to human intelligence.

Humans are required to direct Datamining at every phase. In fact, Human qualitative intelligence such as domain knowledge and field supervision complimenting Mining quantitative intelligence like computational capability will help problem solving. Mining systems have all the computational capability, these needed to be driven in a right direction and this, happens only with the Human experience, metaknowledge and imaginary thinking. When working on the business specifics, it was required adding up business concerns to mining systems, and quickly evaluate mining results. This assistance can largely improve the effectiveness and efficiency of mining actionable knowledge.

## .2 Explain each of the six common data mining tasks.

The most common data mining tasks are,

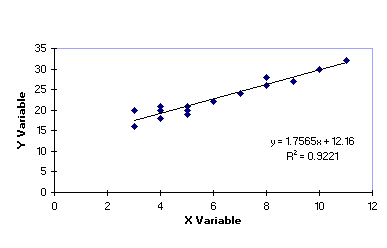
* Description
* Estimation
* Prediction
* Classification
* Clustering
* Associations

1. **DESCRIPTION:** ‘Description’ is one of Data mining tasks. Describing patterns and trends lying in the data often suggest possible explanations. This requires transparency from Data mining models. Which means, the results of the Data mining models should describe clear patterns in the data, which can help in explanation and quickly evaluating results.

One Data Mining method, decision trees provide a clear, human-friendly explanation to their results. In contrast to Neural networks, which are nonlinear and complex models.

Exploratory data analysis helps in accomplishing high quality data description. Which is graphical method of exploring patterns and trends in data.

1. **ESTIMATION:**  Approximating a numeric value of target variable using a set of numeric/ categorical predictors. A specific model is built up using a set of predictor variables (past values of a variable) and try to estimate the value of the new observations.



Let us consider an Regression equation from figure, y=12.16+1.7565x.

This equation is made up using previous data of ‘x’ and ‘y’. using this regression equation, we can estimate the value of y is 12.16 plus 1.7565 times the value of x.

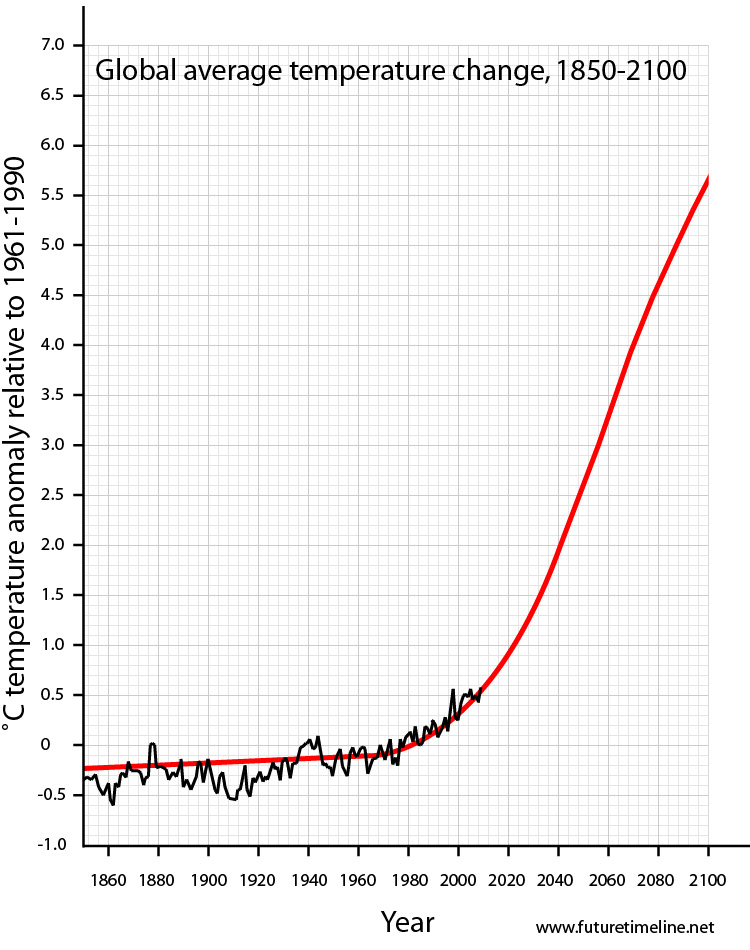
Therefore, we can estimate that Point (x=1.0, y=13.9165) lies on the regression line.

The field of statistical analysis supplies several venerable and widely used estimation methods. These include point estimation and confidence interval estimations, simple linear regression and correlation and multiple regression.

1. **PREDICTION:**

Prediction is very much like estimation methods. Except that the predictive results of this method lie in the future.

An Example of predictive tasks as explained below,



From the above graph, a predictor method is applied to approximate global temperature for the next 90 years.

Any of the methods used for classification and estimations, ca also be used for prediction, under appropriate conditions. These include the traditional statistical methods of point estimation and confidence interval estimations, simple linear regression and correlation, and multiple regression.

1. **CLASSIFICATION:**

Classification is similar to estimation methods, expect that the target variable is categorical instead of numeric. By application of classification methods on independent variables, a categorical target variable is populated.

The algorithm works on the independent variables, learns about the data distribution and categories the data.

For example,

* Classifying the students into a particular track base on their special needs.
* Which drug should be prescribed to a patient aged 63.

1. **CLUSTERING:**

Clustering refers to grouping of data into similar objects. A cluster pulls out all the similar data together. Unlike Classification, there won’t be a target variable in clustering method. The clustering does not try to classify, predict or estimate the value of a target variable.

Clustering method will segment the whole data into relatively homogeneous subgroups, where the similar records with in the cluster are maximized and the dis-similarity is maintained with the other records.

As preliminary step in data mining, Resulting clusters used as input to different technique downstream, such as neural networks.

1. **ASSOCIATION:** The association method for data mining is the job of finding which attributes ‘go together’.

In business world, it is commonly known as affinity analysis or market basket analysis.

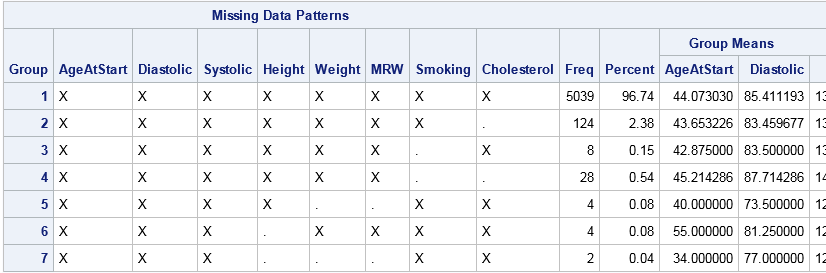
Examples for Association tasks in business include,

* Predicting degradation in telecommunications networks,
* Finding out which items in a supermarket are purchased together, and which items are never purchased together

## .8) Describe 3 methods for handling missing data

We should be careful while handling the missing data, otherwise NULL values in data will certainly affect the resultant information. Mostly we face these things while handling large dataset which usually have large number of fields.

Missing data fields example



One way of handling missed data is omitting the records with NULL value. This may lead to biased subset of the data.

Other handling methods:

1. Replace the missing value with any constant value.
2. Replace the missing value with field mean/mode.
3. Replace with a value generated at random from the observed distribution of the variable.

Replacing the missing value with any user defined constants refers to populating the field with default value depending on the data type of the field. Such as 0 for Integer data field and ‘Missing’ for any textual data type.

Replacing the missing value with their respective means and mode, replacing a field with a value which occurred most number of times. And when replacing with its mean, calculate the mean of the rest of all the values and replace the missing value with the calculated mean value.

Replacing the missing value with any random value generated from observed distribution refers to replacing the NULL value with any randomly generated data depending on the data type. The value being replacing should be selected from the observed distribution of the table.