Computer Science M5 T2 Midterm Exam Short

Data Analysis

Data analysis is a process of turning information into a valuable resource by using the results obtained from the analysis. This will give us a deep understanding of information such as relationships, patterns, and trends. When we analyze something, we always have 3 important questions:

1. What happened? 2. What will happen? 3. What should I/we do?

The answers can be obtained from analyzing 3 types of data:

- Descriptive analytics **- Predictive analytics** **- Prescriptive analytics**

Descriptive analytics – is a fundamental analysis, it gives an overview of data and the relationship between them. Explains what happened in the past so it will help us make choices in the future.

Predictive analytics - is an analysis that helps predict or predicts what is likely to happen in the future. By using the historical data, we can possibly predict what can happen.

Prescriptive analytics - is an analysis to build on the predictions that may happen. We simulate the possible options of a simulation and predicting the outcome of each situation. This way we can suggest which situation would be the best and its possible results.

Descriptive analytics

Is a fundamental analysis, it gives an overview of data and the relationship between them. Explains what happened in the past so it will help us make choices in the future. It uses mathematical calculations and basic statistics such as proportions and percentages. It also measures the central value of the data (median) and distribution of information

A **proportion** is an equation that says that two or more ratios are equal. A **percentage** is a number or ratio that represents a fraction of 100. Its symbol is %. Proportions and percentages make it easier to understand the data because it can be represented with a diagram or a chart.

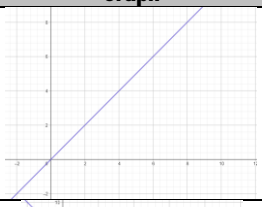
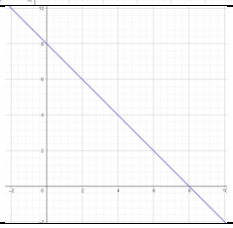
A **median** is a value separating the higher half from the lower half of some data. The median is not the same as average.
Median is mostly used for a small spread of information. **Average** is used when the spread of information is large. The **standard deviation** is used to measure the spread of information and it is always positive. If **the deviation number is low**, that means that the dataset values are **close to each other**. If the **deviation number is high**, that means that the dataset **values are spread over a wider range**.

Dataset relationship

A **data set (or dataset) is a collection of data**. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Data sets can also consist of a collection of documents or files.
Analysis of the relationship between two datasets shows the **direction** of the relationship and the **degree** of the relationship

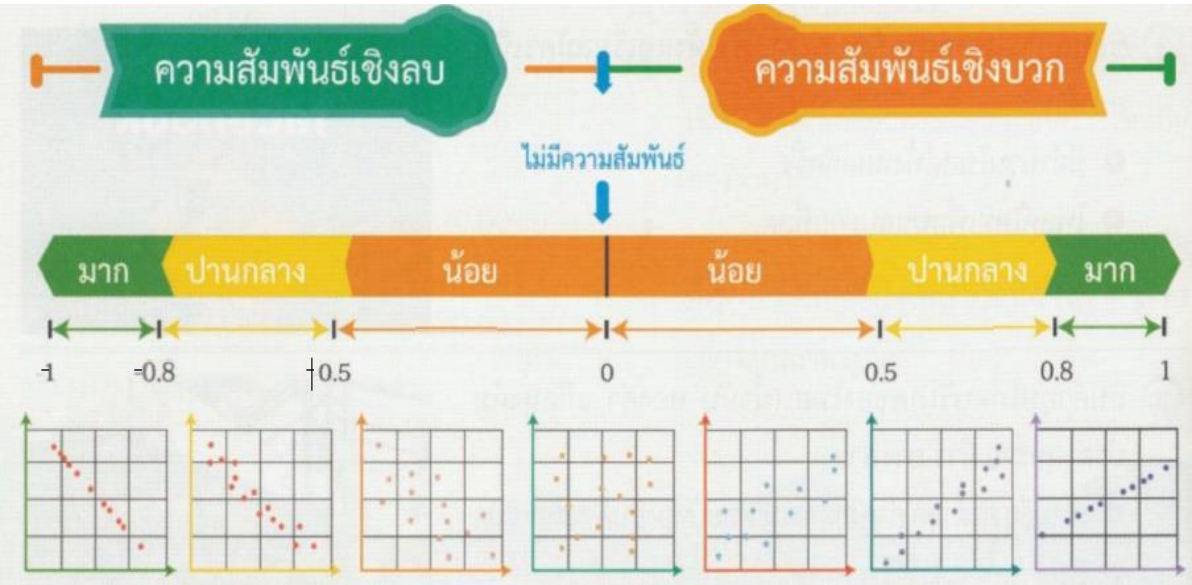
Direction of the relationship means in which way is the data relationship going. It is also referred as correlation. We have two types of relationship:

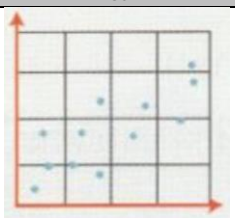
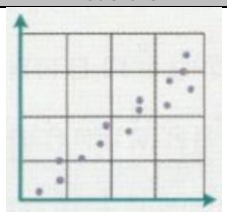
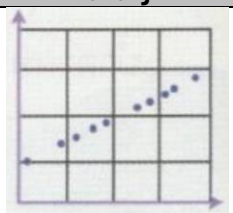
- Positive - Negative

Relationship	Explanation	Graph
Positive	A positive direction means that both interests increase or decrease in the same direction. For example, the value of x is increasing ↑ and the value of y is also increasing ↑	
Negative	A negative direction means that both interests increase or decrease in opposite direction. For example, if the value of x increases ↑ then the value of y will decrease ↓	

Degree of the relationship shows us how much are the data sets related to each other. There are 3 degrees of relationship:

- **Strong** – the data is aligned with each other
- **Moderate** – the data isn't aligned perfectly
- **Weak** – the data is not aligned at all, the data is placed everywhere



Direction/Degree	Weak	Moderate	Strong
Positive			
Negative