STA 3180 Statistical Modelling: Big Data

STA 3180 Statistical Modelling: Lecture Notes on Big Data

Big Data is a term used to describe large and complex datasets that are difficult to process using traditional data processing techniques. It is characterized by its volume, velocity, variety, veracity, and value. Big Data is increasingly being used in various industries, such as healthcare, finance, and retail, to gain insights into customer behaviour, market trends, and other business-related information.

Key Concepts:

1. Volume: Big Data is characterized by its large size, which can range from terabytes to petabytes.

2. Velocity: Big Data is characterized by its high speed of data generation and collection.

3. Variety: Big Data is characterized by its diverse types of data, such as structured, semi-structured, and unstructured data.

4. Veracity: Big Data is characterized by its uncertain quality, which can lead to inaccurate results.

5. Value: Big Data is characterized by its potential to provide valuable insights into customer behaviour, market trends, and other business-related information.

Definitions:

1. Big Data: Large and complex datasets that are difficult to process using traditional data processing techniques.

2. Structured Data: Data that is organized into a fixed format, such as tables or spreadsheets.

3. Semi-structured Data: Data that is organized into a flexible format, such as XML or JSON.

4. Unstructured Data: Data that is not organized into any specific format, such as text, images, or audio.

Coding Example:

```
// Start of Code
import pandas as pd
# Read in the dataset
df = pd.read_csv('dataset.csv')
# Print the first 5 rows of the dataset
print(df.head())
# End of Code
```

Practice Multiple Choice Questions:

Q1. What is Big Data?

A. Large and complex datasets that are difficult to process using traditional data processing techniques.

B. Small datasets that can be easily processed using traditional data processing techniques.

C. Structured datasets that are organized into a fixed format.

D. Unstructured datasets that are not organized into any specific format.

Answer: A. Large and complex datasets that are difficult to process using traditional data processing techniques.