Unit - 4

1) Bivariate analysis:

* The term Bivariate analysis refers to analysis of two variables.

* It is a methodical statistical technique applied to pair of variable of data to determine relationship.

Common ways to Perform Bivariate.

1) Scatter plots.

Scatter plots gives an

idea of the patterns that can be formed using two variables

Correlation Coefficient:

* The coefficient helps to know if the data's are correleated.

* When the coefficient relation is zero then this means that the variable are not related.

* If the correlation coefficie is a positive or negative 1. then this means that the variables are prefectly correlat

linear regression:

* This uses a wide range of tools to determine how the data is related.

* It represented by line or curve.

2) Percentage tables:

* percentage tables are also known as frequency tables or relative frequency tables.

* It is a common tool in data analysis to summarize and present the categorical data

* percentage table provides a clear picture of proportion

* It also helps to understand the distribution of different categories.

* Here is a step by step guide to create and interpret percentage tables.

1) Define your Categories:

* Identify the categorical variable you want to analyze.

* categorical variables consist of distinct groups or categories such as age groups, gender, productstype

2. Collect and organize data:

Collect your date and organize in table format

3. Create a Frequency table:

* Create a frequency table by counting the number of occurrences of each category in dataset

for example.

Category	Frequency
Male	15
Female	12
Others	2

4. Calculate Percentages:

* Add a column to your frequency table to calculate the percentage of each category.

* Divide the frequency of each category by the total number of observations and multiply by 100.

for example:

Category	Frequency	Percentage.
Male	15	51.1.
female	12	41.1.
others	2	6:1.

5). Interpret the Result:

* Analyze the percentage table. to understand the distribution of categories

3. Handling Several Batches:

* Many analytics applications require frequent batch processing.

* Batch processing allows to process data in batches at varying interval

* Batch systems must be built to scale for all sizes of data.

* Data received from batch processing is refferred as big data or large data.

* Batch processing is a technique of processing large amount data is broken down into smaller chunks for bebugging efficiency.

Reduce memory usage:

* pandas automatically reduces the memory usage by optimizing data types.

Splits data into chunks.

* when data is too large to fit into memory, pandas can be used.

* pandas split data into smaller chunks, which makes a easy deal to debugging and filtering the dataset.

* Chunking can be used from intial stage of exploratory analysis.

Benefits

* Speed and Low costs.

* offline feature.
* Efficiency
* Simplicity
* Improved Data Quality