Disclaimer

- This presentation is purely for academic purpose and does not carry any commercial value.
- All images and photos used in this presentation are property of respective image holder(s) and due credit is provided to them. Images are used only for indicative purpose and does not carry any other meaning.
- All information and data in this slide are collected from open domain.

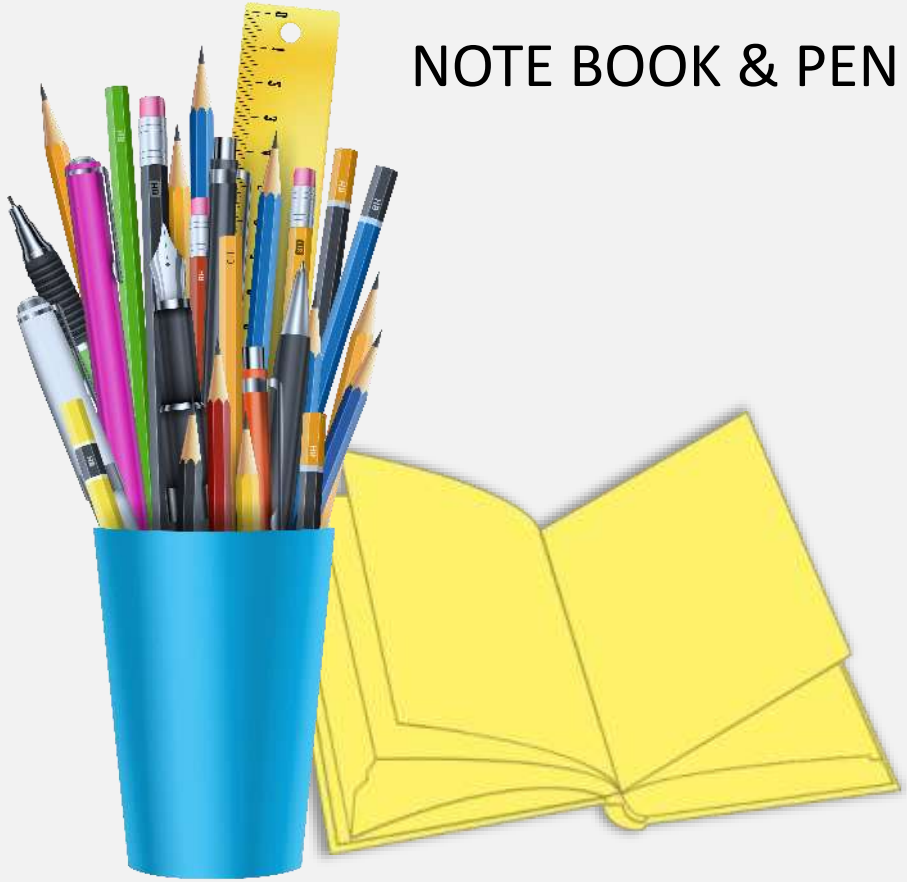
MANISH GODSE, Ph.D.(IIT Bombay)

Request & Instructions

PLEASE OPEN

CALCULATOR

NOTE BOOK & PEN



LAPTOP OR DESKTOP,
IF YOU HAVE.

PLEASE FOLLOW THIS

SILENCE

MUTE MIC

RAISE HAND

NO CHAT

SILENT MODE

DESCRIPTIVE STATISTICS

BOOKS & REFERENCES

- NO

Table of Contents

STEPS TO PERFORM

1

WHAT IS DESCRIPTIVE STATISTICS?

- Descriptive statistics is used to summarize data in more meaningful way so that it allows for simpler interpretation of data and helps generate insight into the information contained in the data.
- Choosing which summary statistics are appropriate depend on the type of variable being examined.
- Different statistics should be used for ratio, ordinal, and nominal data.
- Raw data is used in descriptive statistics.
- Descriptive statistics is important to establish validity that sample is representing correct population.

What is Examined in Descriptive Statistics?

LOCATION is called central tendency. It is a measure of the values of the data. Measures of location include mean, median etc.

VARIATION is also called dispersion. It is a measure of how far the data points lie from one another. Common statistics include standard deviation and coefficient of variation. For data that aren't normally-distributed, percentiles or the interquartile range might be used.

SHAPE refers to the distribution of values. The best tools to evaluate the shape of data are histograms and related plots. Statistics include skewness and kurtosis, though they are less useful than visual inspection. We can describe data shape as normally-distributed, log-normal, uniform, skewed, bi-modal, and others.

DESCRIPTIVE STATISTICS

MEASURE	STATISTICS METHOD
Size of Data	Number of rows and columns
Data Types	Check data type of each variable or attribute
Measure of Location (or Center)	Mean, Median, Mode, Geometric Mean, Harmonic Mean
Measure of Dispersion (or Dispersion)	Minimum Value, Maximum Value, Range of Data, Variance, Standard Deviation, Coefficient of variance
Measure of Distribution (or Shape)	Percentile, Quantile, Quartile, Inter-Quartile Range, z-score, skewness & kurtosis, Box plot
Measure of Association	Scatter chart, Covariance, Correlation coefficient

PLOTTING DISTRIBUTION

Number of data	Data type	Graph
Univariate data	Discrete data	Histogram
	Continuous data	Polygon
Bivariate data		Scatter plot

STEPS TO PERFORM IN PYTHON

- Import Libraries
- Read data from CSV file
- Save data as DataFrame
- Dimension of Data
  - Get names of columns
  - Get number of rows and columns
- Check Data Types
- Sample of Data
  - View sample data
  - Get top 5 rows
  - Get top bottom 5 rows

- Summary of Data
- Standard Deviation
- Skewness and Kurtosis
- Correlations
- Plot/Chart

EXAMPLE - BOSTON HOUSING

2

GET DATA

```
# import pandas library
import pandas as pd

# Get the file path
sales_data = "E:/Courses/Machine_Learning_Python/data/boston-housing/boston_housing.csv"

# Read the file and save as DataFrame
sdf = pd.read_csv(sales_data, header=0)
print('Original DataFrame of Input File\n', sdf)
```

EXPLORE DATA

```
# Get shape (number of rows and columns) of the Boston Data
print("\nShape of Boston Dataset = ", sdf.shape)
# Get the names of features of the Boston Dataset
print("\nFeature names of Boston Dataset = \n", sdf.columns)
# Get heads of the Boston Dataset
print("\nPrint top records of the Boston Dataset = \n", sdf.head())
# Print datatype
print("\nPrint data types of the Boston Dataset = \n", sdf.dtypes)
# View top 5 data
print("\nPrint top 10 rows of the Boston Dataset = \n", sdf.head(10))
# View bottom 5 data
print("\nPrint bottom 10 rows of the Boston Dataset = \n", sdf.tail(10))
```

CHECK DUPLICATE & MISSING DATA

```
# Get count of duplicate rows
print('\nCount of duplicate records = ',sdf.duplicated().sum())

# Get count of missing values
print('Count of missing value in each column =')
print(sdf.isnull().sum())
```

PLOT HISTOGRAM

```
# Import seaborn library
import seaborn as sns

# Set histogram
sns.set()
sdf.hist(figsize=(10,7), color='blue')

# Import matplotlib library
import matplotlib.pyplot as plt

#Show histogram
plt.show()
```

DEVELOP PAIR PLOT

```
# Import seaborn library
import seaborn as sns
sns.pairplot(sdf)

# The pairs plot builds on two basic figures, the histogram and the scatter plot.
# The histogram on the diagonal allows us to see the distribution of a single variable
while the scatter plots on the upper and lower triangles show the relationship (or lack
thereof) between two variables.

# Get DISTRIBUTION PLOT for a single variable - 'PRICE'
import seaborn as sns
sns.distplot(sdf['PRICE'])
```

CORRELATION OF DATA AND HEATMAP

```
# Get correlation of a DataFrame
print(sdf.corr())
# import seaborn library
import seaborn as sns
# Create Heatmap of correlation using seaborn library
sns.heatmap(sdf.corr())
# import matplotlib library
import matplotlib.pyplot as plt
# Show Heatmap of correlation using matplotlib library
plt.show()
```

DESCRIPTIVE STATISTICS

```
# View descriptive statistics of data
sdf.describe()
```

QUESTION AND ANSWERS