UMBER NOREEN

Data Analytics
Week

WEEK

- Descriptive Statistics
  - Data Summarization
    - Mode
    - Median
    - Mean
  - Measure of Dispersion/Variability
    - Range
    - Variance
    - Standard Deviation

DATA

A set of data is collection of observed values representing one or more characteristics of some objects or units.

Data is consist of many attributes.

POPULATION

A population is a data set representing the entire entities of interest.

For Statistical learning, it is very important to define the population that we intend to study very carefully.

SAMPLE

A Sample is a data set consisting of a population.

A sample is obtained in such a way as to be represented of the population.

STATISTICS

A Staistics is a quantity calculated from data that describes a particular characteristics of a sample.

STATISTICAL INFERENCE

A process of using sample statistics to make decisions about population.

DATA SUMMARIZATION

- To Identify the typical characteristics of data (to have an overall picture)
- To identify which data should be treated as noise or outliers.

- It is classified broadly in two categories:

  - Measure of Location

  - Measure of Dispersion

MEASURE OF LOCATION

- It is also called as Measuring the Central Tendency
- A function of a sample values that summarizes the location information into a single number is known as measure of location.

- The Popular Measure of Locations are:

  - Mode
  - Mean
  - Median
  - Midrange

MODE

The mode is the most frequently occurring value in a set of data.

- One mode → Unimodal
- Two modes → Bimodal
- More than two modes → Multimodal

MODE

Offer Prices for the 20 Largest U.S. Initial Public Offerings in a Recent Year

$14.25	$19.00	$11.00	$28.00
24.00	23.00	43.25	19.00
27.00	25.00	15.00	7.00
34.22	15.50	15.00	22.00
19.00	19.00	27.00	21.00

MEDIAN

- The median is the middle value in an ordered array of numbers.

- STEP 1. Arrange the observations in an ordered data array.
- STEP 2.For an odd number of terms, find the middle term of the ordered array. It is the median.
- STEP 3.For an even number of terms, find the average of the middle two terms. This average is the median.

MEDIAN

- Suppose a business researcher wants to determine the median for the following numbers.

15 11 14 3 21 17 22 16 19 16 5 7 19 8 9 20 4

Median?

15 11 14 3 21 17 16 19 16 5 7 19 8 9 20 4

MEAN

- The arithmetic mean is the average of a group of numbers
- Population Mean:

$$\mu = \frac{\Sigma x}{N} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}$$

- Sample mean:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

MEAN

- Compute the mean of the following numbers
- 17.3 44.5 31.6 40.0 52.8 38.8 30.1. 78.5
- 7 -2 5 9 0 -3 -6 -7 -4 -5 2 -8

MEAN

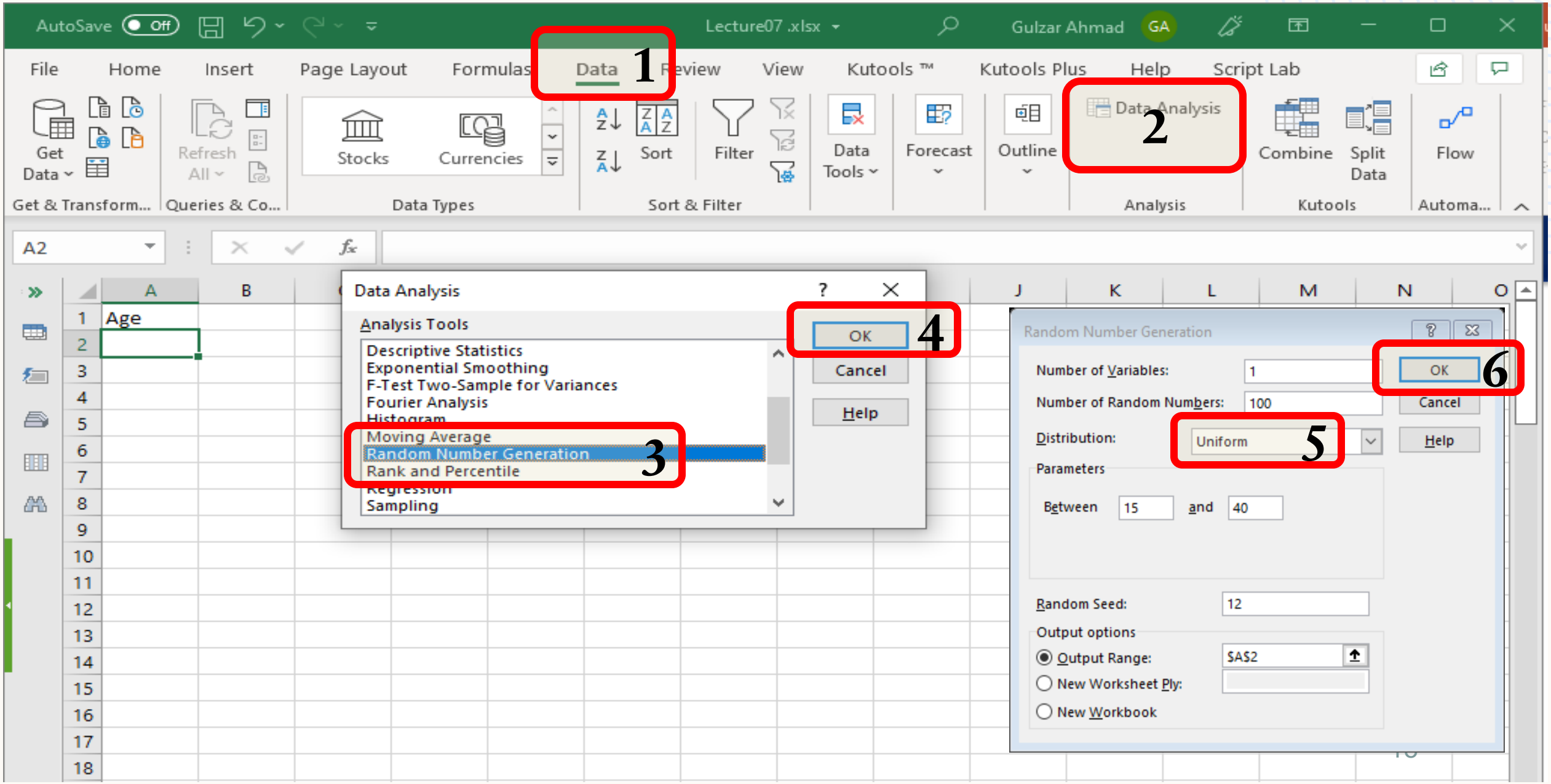
- The number of U.S. cars in service by top car rental companies in a recent year according to Auto Rental News follows.

- Calculate the mode, median, mean

Company	Number of Cars in Service
Enterprise	643,000
Hertz	327,000
National/Alamo	233,000
Avis	204,000
Dollar/Thrifty	167,000
Budget	144,000
Advantage	20,000
U-Save	12,000
Payless	10,000
ACE	9,000
Fox	9,000
Rent-A-Wreck	7,000
Triangle	6,000

USING PYTHON

- # Aggregating Functions in NUMPY
- a = np.array([13,23, 20, 30, 27, 10])
- np.sum(a) # Sum of all the value in the Array
- np.min(a) # Minimum Value in the Array
- np.max(a) # Maximum Value in the Array
- np.mean(a) # Mean Value of the Array
- np.median(a) # Median Value of the Array
- np.average(a) # Gives Average of the array values
- np.var(a) # Gives Variance
- np.std(a) # Gives Standard Deviation
- Range = max(a) – min(a) # Gives range
- Cv = (np.std(a)/ np.mean(a)) * 100 # Coefficient of Variance
- np.size(a) # Number of values/elements in an arrays


AutoSave Off
Lecture07 .xlsx
Gulzar Ahmad GA
File Home Insert Page Layout Formulas Data Review View Kutools ™ Kutools Plus Help Script Lab
1
Get Data
Refresh All
Stocks
Currencies
Sort
Filter
Data Tools
Forecast
Outline
Data Analysis
2
Combine
Split Data
Flow
Get & Transform...
Queries & Co...
Data Types
Sort & Filter
Analysis
Kutools
Automa...
A2
Age
Data Analysis
Analysis Tools
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
3
OK
Cancel
Help
4
Random Number Generation
Number of Variables: 1
Number of Random Numbers: 100
Distribution: Uniform
5
Parameters
Between 15 and 40
Random Seed: 12
Output options
Output Range: A2
New Worksheet Ply:
New Workbook
OK
Cancel
Help
6

USING EXCEL

- =AVERAGE(DataValues)
- =MEDIAN(DataValues)
- =MODE.SNGL(DataValues)

- E.g.
- =AVERAGE(A1:A13)
- =MEDIAN(A1:A13)
- =MODE.SNGL(A1:A13)