Data Analytics: UE18CS312
Question Bank

Unit-1: Exploratory Data Analysis and Visualization

Sl.No Questions

1. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30,33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
(a) What is the mean of the data?What is the median?
(b) What is the mode of the data? Comment on the data’s modality (i.e., bimodal, trimodal, etc.).
(c) What is the midrange of the data?
(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
(e) Give the five-number summary of the data.
(f) Show a boxplot of the data.
(g) How is a quantile–quantile plot different from a quantile plot?

2. Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows:

Age	Frequency
1-5	200
6-15	450
16-20	300
21-50	1500
51-80	700
81-110	44

Compute an approximate median value for the data.

3. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

Age	2	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61

	3																	
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	24.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median, and standard deviation of age and %fat.
(b) Draw the boxplots for age and %fat.
(c) Draw a scatter plot and a q-q plot based on these two variables.

4. Briefly outline how to compute the dissimilarity between objects described by the following:
(a) Nominal attributes
(b) Asymmetric binary attributes
(c) Numeric attributes
(d) Term-frequency vectors

5. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):
(a) Compute the Euclidean distance between the two objects.
(b) Compute the Manhattan distance between the two objects.
(c) Compute the Minkowski distance between the two objects, using q D 3.
(d) Compute the supremum distance between the two objects.

6. The median is one of the most important holistic measures in data analysis. Propose several methods for median approximation. Analyze their respective complexity under different parameter settings and decide to what extent the real value can be approximated. Moreover, suggest a heuristic strategy to balance between accuracy and complexity and then apply it to all methods you have given.

7. It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation. Suppose we have the following 2-D data set:

	A1	A2
x1	1.5	1.7
x2	2	1.9
x3	1.6	1.8
x4	1.2	1.5
x5	1.5	1.0

(a) Consider the data as 2-D data points. Given a new data point, x = .1.4, 1.6/ as a query, rank

the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.
(b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

8. Data quality can be assessed in terms of several issues, including accuracy, completeness, and consistency. For each of the above three issues, discuss how data quality assessment can depend on the intended use of the data, giving examples. Propose two other dimensions of data quality.

9. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

10. Question no. 1 gave the following data (in increasing order) for the attribute age: 13, 15,16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46,52, 70.
(a) Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.
(b) How might you determine outliers in the data?
(c) What other methods are there for data smoothing?

11. What are the value ranges of the following normalization methods?
(a) min-max normalization
(b) z-score normalization
(c) z-score normalization using the mean absolute deviation instead of standard deviation
(d) normalization by decimal scaling

12. Use these methods to normalize the following group of data:
200, 300, 400, 600,1000
(a) min-max normalization by setting min D 0 and max D 1
(b) z-score normalization
(c) z-score normalization using the mean absolute deviation instead of standard deviation
(d) normalization by decimal scaling

13. Using the data for age given in question no.1, answer the following:
(a) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].
(b) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.
(c) Use normalization by decimal scaling to transform the value 35 for age.
(d) Comment on which method you would prefer to use for the given data, giving reasons as to why.

14. Using the data for age and body fat given in Question no.3, answer the following:
(a) Normalize the two attributes based on z-score normalization.
(b) Calculate the correlation coefficient (Pearson’s product moment coefficient). Are these two attributes positively or negatively correlated? Compute their covariance.

15. Suppose a group of 12 sales price records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.
Partition them into three bins by each of the following methods:
(a) equal-frequency (equal-depth) partitioning
(b) equal-width partitioning
(c) clustering

16. Use a flowchart to summarize the following procedures for attribute subset selection:
(a) stepwise forward selection
(b) stepwise backward elimination
(c) a combination of forward selection and backward elimination

17. Using the data for age given in Question no.3,
(a) Plot an equal-width histogram of width 10.
(b) Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR, cluster sampling, and stratified sampling. Use samples of size 5 and the strata “youth,” “middle-aged,” and “senior.”

18. The daily football at a retail store in Bangalore over the last 30 days is shown in Table 1. calculate the Mean, Median , Mode and Standard Deviation.

Table 1.Footfall data

232	277	261	173	283	197	251	212	213	213
229	164	219	196	186	247	244	269	216	272
252	314	161	165	221	260	219	290	225	251

19. For the data in Table 1, calculate the skewness and kurtosis. what can you infer from the skewness and kurtosis of the football data?

20. For the data in Table 1, calculate the values of first quartile and third quartile. Are there any outliers in the data?

21. The Bank of Kala Bakra(BKB) situated in Bakrapur, India receives several applications for home loan and home improvement loan. The description of the data captured in 'know your customer'(KYC) document listed below.

i. Customer ID
ii. Type of loan (2 types: Home Loan and Home Improvement Loan)
iii. Gender (Male, Female)
iv. Marital Status (Married, Single and Others)
v. Accommodation type (Family Other, Company Provided, Owned, Rented)

vi. Number of years in the current address

vii. Number of years in the current job

viii. Monthly salary in Indian rupee

ix. Balance in savings account (in Indian Rupees)

x. Loan amount requested (In Indian Rupees)

xi. Term (Loan term in months)

xii. Down payment (In Indian Rupees)

xiii. Equal Monthly Installment (EMI) affordable

a. Develop appropriate charts for the variables. What insights can be obtained based on the charts?

b. Calculate the mean, median, mode, variance, standard deviation, skewness and kurtosis of variables monthly salary and balance in saving account.

c. Use box plot to check whether there are outliers among variables loan amount requested, down payment, and EMI.

d. Which variable among continuous variables have high skewness?

22. The cumulative grade point average(CGPA) of 40 students are shown in Table 2: CGPA of students

3.36	1.56	1.48	1.43	2.64	1.48	2.77	2.20	1.38	2.84
1.88	1.83	1.87	1.95	3.43	1.28	3.67	2.23	1.71	1.68
2.57	3.74	1.98	1.66	1.66	2.96	1.77	1.62	2.74	3.35
1.80	2.86	3.28	1.14	1.98	2.96	3.75	1.89	2.16	2.07

a.Calculate the mean, median, and mode. Calculate the standard deviation.

b.Calculate the 90th and 95th percentile of CGPA

C. Calculate the inter quartile range (IQR)

D. The dean of the school believes that the CGPA is a right tailed distribution. Is there an evidence to support deans belief?

E. Create a histogram for the data, what should be the ideal number of bins in the histogram.

23. Value of insurance claims at an insurance company has mean value of INR 7200 and standard deviation of 200. Comment on the proportion of claims with values between INR 6900 and 7500.

24. Demand for a spare parts sold by a capital equipment manufacturer is shown in

Table 4:Demand for spare parts

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
65	106	55	98	80	84	105	111	103	137

25. Share Khan and Sons (SKS) is an investment advisory company. SKS has identified top 50 shares and its values rounded to nearest rupees are shown in Table .Value of shares in rupees.

600	349	292	247	216	411	233	364	419	505
474	541	790	293	362	470	349	429	565	309
453	419	354	273	533	235	467	569	590	347
413	541	318	545	256	247	474	597	522	535
483	573	345	568	260	288	50	248	466	417

a. Plot a histogram for the data. What insights can you gain from the histogram?
b. Plot a box plot and identify if there are any outliers
c. Is the distribution of share price mesokurtic? Respond using an appropriate measure.
d. If the value of share is 600, calculate its percentile value.

26.

27. On 8 November 2016, Indian government demonetized 500 and 1000 rupees notes and allowed the citizens to deposit the old notes in the banks. The average amount deposited by customers at a bank in Bannerghatta road is 17500 and the corresponding standard deviation is 4500. If 156 customers deposited money on 11 November 2016.
(a) Calculate the probability that the total deposits exceed INR 3 million.
(b) What is the probability that the total deposited amount is between INR 2.5 million and INR 3.5 million?

28. Average time to churn of a telecom customer is estimated as 300 days from a sample of 10000 customers. What is the probability of observing this sample mean of at least 300 days

from a population in which the mean time to churn is 280 days and standard deviation is 72 days?

29. The proportion of defaults in mortgage loan is estimated as 8% in the population. In a sample of 1000 mortgage loans, what is the probability that the proportion of defaults will exceed 10%?

30. The cost to company (CTC) of 50 IT professionals measured in lakhs of rupees is shown in Table 1.

TABLE 1. CTC (in lakhs of rupees)

21.38	12.24	29.06	12.37	8.48	18.76	23.8	28.48	9.56	35.94
28.76	30.76	37.67	34.15	32.53	26.64	24.25	39.66	8.98	26.17
40.54	27.66	18.83	12.87	22.12	28.07	27.15	12.06	5.66	8.44
4.85	11.72	15.18	6.44	28.94	17.71	31.5	26.91	33.93	14.5
38.14	30.87	27.29	6.77	18.43	28.9	22.33	31.41	37.03	32.6

(a) Draw a histogram. Comment on the distribution of CTC using skewness and kurtosis.

(b) Generate 500 random samples of size 10 and plot the histogram of sampling distribution.

(c) What is the mean and standard deviation of the sampling distribution obtained in (b)?

How far is this mean from the mean CTC of values provided in Table 1?

31. The amount of time that a vehicle has to wait at a traffic signal in the city of Bangalore is uniformly distributed between 3 and 16 minutes. Use method of moments to estimate the average waiting time and standard deviation of waiting time.

32. Use maximum likelihood estimate to find the scale and shape parameters of a two-parameter Weibull distribution with probability density function

$$f(x) = \beta/n(x/n)^{\beta-1}\ e^{-(x/n)\ \beta}$$

where h is the scale parameter and b is the shape parameter.

33. Call duration of calls made by customers of a telephone company follows an exponential distribution with mean 320 seconds and standard deviation 80 seconds. Calculate the probability that the average call duration of a random sample of 250 calls will exceed 300 seconds.

34. Waiting time at a bank follows a normal distribution with mean 16 minutes and standard deviation 4 minutes. Calculate the sample size required to estimate the mean at a confidence of 95% and maximum error in estimation of 2 minutes.

35. According to the government sources, 1% of 1000 rupees currency notes are counterfeit notes. If one would like to estimate the percentage of counterfeit notes in circulation within an error range of 0.001, calculate the sample size required at a = 0.01.

36. Table 2 shows time to failure of air conditioners (ACs) sold by a company measured in days since sale. The time to failure is assumed to follow an exponential distribution. Calculate the mean time between failure (1/l) using maximum likelihood estimate. If the warranty period is 365 days, calculate the proportion of ACs likely to fail during the warranty period.

TABLE 2 Time to failure (measured in days) of air conditioners

86	554	318	366	1180	175	74	276	653	438
284	161	32	342	470	701	314	989	586	3151
295	3999	1790	116	272	176	80	215	1770	733
1751	1809	142	888	200	501	237	304	1563	252
106	372	1097	133	145	69	201	3070	957	111

37. On 8 November 2016, Indian government demonetized 500 and 1000 rupees notes and allowed the citizens to deposit the old notes in the banks. The average amount deposited by customers at a bank in Bannerghatta road is 17500 and the corresponding standard deviation is 4500. If 156 customers deposited money on 11 November 2016.

(a) Calculate the probability that the total deposits exceed INR 3 million.

(b) What is the probability that the total deposited amount is between INR 2.5 million and INR 3.5 million?