1. Which of the following would extract all the rows of the first 3 columns in a given numpy 2D array ‘a’?
2. a[ : , :3] B) a[ : , :2] C) a[( : , :2)]  D) a[( : , :3)]
3. Consider an (11, 12) shape array. What is the index (x,y) of the 100th element?

Note: For counting the elements go row-wise. For example, the array,

[[1, 5, 9],

  [3, 0, 2]]

the 5th element would be '0'

1. Swap the 0th and 2nd rows of the 2-D NumPy array given below

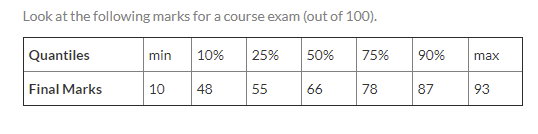
a = np.array ([[4, 3, 1], [5, 7, 0], [9, 9, 3], [8, 2, 4]])

**For Q4 – Q6, download and refer the cricket.ipynb file and answer the questions**

1. Which of the following commands can you use to extract the top 4 teams in the dataset ‘ipl18’ with just the ‘Team’ and ‘Points’ column?
2. ipl18.loc [0:4, [‘Team’, ‘Points’]]
3. ipl18.loc [1:4, [‘Team’, ‘Points’]]
4. ipl18.loc [0:3, [‘Team’, ‘Points’]]
5. ipl18.loc [1:3, [‘Team’, ‘Points’]]
6. Suppose in ‘ipl18’, you want to filter out the teams that have an NRR greater than zero, and for which the ‘For’ score exceeds the ‘Against’ score, i.e. both the conditions should be satisfied. Which teams will be left after you perform the above filtration? (Run the commands on the Python Notebook provided, rather than performing a manual calculation)
7. CSK, MI
8. SRH, CSK, MI, RCB
9. SRH, CSK, RCB
10. SRH, CSK, MI
11. If all the stats are taken for both ‘ipl17’ and ‘ipl18’, which team with its total points greater than 25 will have the highest win percentage?
12. KKR
13. RPS
14. SRH
15. CSK

**Refer “ODI\_score.csv” for answering Q7- Q9.**

1. Plot a bar chart showing runs scored on the x-axis and frequency/count on the y-axis. In which bucket has he scored runs the most often?
2. 0-10 runs
3. 10-20 runs
4. 20-30 runs
5. 30-40 runs
6. Plot a histogram of the number of 4s hit by the player. What is the most common value of the variable X, where X represents the number of 4s hit by him?
7. 0
8. 1
9. 2
10. 3
11. In the 'Runs' column of the given dataset, identify all the data quality issues. Choose all the correct options
12. Some values are 'TDNB'
13. Some Values are ‘DNB’
14. Some values contain a "\*", such as 8\*
15. Some values are extremely large and represent an anomaly (may be a data entry issue)



Which of the following statements is FALSE?

1. About 1/4 of the class received a score of 55 or less
2. About 3/4 of the class received a score of 78 or less
3. About 50% of the class received grades between 55 and 78
4. About 1/30 of the class received a score of 48 or less

**Refer popularity.csv for answering Q11- Q17. Use boxplots for Q11-Q14**

1. The variation in the number of shares is larger for:.
2. Weekdays
3. Weekends
4. Say you want to compare the number of shares of articles across weekdays. Select the correct statement
5. There is a significant difference in the spread of the number of shares across weekdays
6. Articles published on Mondays get significantly more popular than those on other days
7. There is no significant difference in either the median or the spread of the number of shares across weekdays
8. Create boxplots for shares vs the channel types and answer which types of articles are most likely to reach 2500 shares?
9. Technology
10. Lifestyle
11. Social Media
12. World
13. What is the mode of the num\_keywords variable in the News Popularity data set?
14. 7
15. 8
16. 9
17. 10
18. What is the average number of times the articles in the data set were shared, i.e what is the mean of the shares?
19. 2427
20. 7322
21. 3395
22. 5000
23. What is the median value of the shares?
24. 7322
25. 3598
26. 1500
27. 1400
28. Choose the correct statement:
29. The mean is comparable to the median
30. The maximum number of share is 843300
31. The mean number of shares is lower than the median

**Refer “achievements.csv” for answering Q18 – Q20**

1. What is the impact of the variable 'Watch.TV' on Science marks?
2. Students who watch TV daily get significantly better marks than others
3. Students who watch TV once a week get significantly better marks than others
4. Students who never watch TV get significantly better marks than others
5. The Watch.TV variable doesn’t distinguish the students significantly in terms of Science marks
6. What is the impact of the father’s education on the student’s Maths marks?
7. Students whose fathers have a degree and above get better marks on an average
8. Students whose fathers are uneducated get better marks on an average
9. The father’s education does not have a significant impact on the student’s average maths marks
10. What is the impact of 'Play.Games' on Reading marks?
11. Students who play games daily score high in Reading on an average
12. Students who play games once a week score high in Reading on an average
13. Students who never play games score high in Reading on an average
14. Students who plays games once a month score high in Reading on an average
15. a) The painkiller drug needs to have a time of effect of at most 200 seconds to be considered as having done a satisfactory job. Given the same sample data (size, mean, and standard deviation) of the previous question, test the claim that the newer batch produces a satisfactory result and passes the quality assurance test. Utilize 2 hypothesis testing methods to make your decision. Take the significance level at 5 %. Clearly specify the hypotheses, the calculated test statistics, and the final decision that should be made for each method.

b) You know that two types of errors can occur during hypothesis testing — namely Type-I and Type-II errors — whose probabilities are denoted by α and β respectively. For the current sample conditions (sample size, mean, and standard deviation), the value of α and β come out to be 0.05 and 0.45 respectively

Now, a different sampling procedure(with different sample size, mean, and standard deviation) is proposed so that when the same hypothesis test is conducted, the values of α and β are controlled at 0.15 each. Explain under what conditions would either method be more preferred than the other, i.e. give an example of a situation where conducting a hypothesis test having α and β as 0.05 and 0.45 respectively would be preferred over having them both at 0.15. Similarly, give an example for the reverse scenario - a situation where conducting the hypothesis test with both α and β values fixed at 0.15 would be preferred over having them at 0.05 and 0.45 respectively. Also, provide suitable reasons for your choice (Assume that only the values of α and β as mentioned above are provided to you and no other information is available)

1. Give 2 examples of each of the following data types:
2. Continuous data
3. Discrete Data
4. Ordinal Data
5. Nominal Data
6. What are the type-1 error and type-2 error?

**Answer Q24- Q26 in the assignment notebook named “Movie\_Assignment.ipynb”**

Q1- Q20 [2x20 = 40 Marks]

Q21 [6 Marks]

Q22.[2 Marks]

Q23.[2 Marks]

Q24 [ 2x2 = 4 Marks]

Q25 [ 3x6 = 18 Marks]

Q26 [ 4x7 = 28 Marks]