भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Certificate Program in Business Analytics and Intelligence (BATCH 9)

Module 1: Data Visualization and Data Interpretation Assignment

Instructions

1. This is a **take-home** assignment. You are free to discuss the assignment questions with your classmates. However, you are not allowed to copy the answers from other students.
2. Answer all questions; there are 10 questions in the assignment. You have to download the following case studies (and the corresponding datasets) from Moodle:
    1. Analytics Empowering Agriculture: Jayalakshmi Agro Tech
    2. A Dean's Dilemma: Selection of students for MBA program and the corresponding dataset.
3. **Show all work and give adequate explanations to get credit. The mathematical equations of the solutions should be clearly mentioned and Outputs must be interpreted and clearly explained. Software outputs without mathematical explanation will NOT be given credit.**
4. Codes run in tools need not be shared. Excel/other software outputs alone shall not be treated as complete answers
5. **Encircle or underline your final answer for each part.**
6. Use significance of 0.5 ($\alpha = 0.05$) where ever required.
7. **Completed assignment should be submitted in Moodle by 16th August 2018 23.55 Hours. Assignments will not be accepted after 16th August 2018, the grades of such students will be marked Incomplete (I) in the grade sheet.**

Question 1 (5 Points)

The cumulative grade point average (CGPA) of 40 students is shown in Table 1.

Table 1 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 90^{th} and 95^{th} percentile of CGPA

(c) Calculate the inter quartile range (IQR)

(d) The Dean of the school believes that more students are towards right tail of the distribution, is there evidence to support dean's belief?

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

Question 2 (10 Points)

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 is listed below (Data file: BKB.Xls):

2. Customer ID
3. Type of loan (2 types: Home Loan and Home Improvement Loan)
4. Gender (Male, Female)

5. Marital Status (Married, Single and Others)
6. Accommodation type (Family Other, Company Provided, Owned, Rented)
7. Number of years in the current address
8. Number of years in the current job
9. Monthly salary in Indian rupees
10. Balance in savings account (Indian Rupees)
11. Loan amount requested (in Indian Rupees)
12. Term (loan term in months)
13. Down payment (in Indian rupees)
14. Equal Monthly Installment (EMI) affordable

(a) Develop appropriate charts for the variables. What insights can be obtained based on the charts? (3 points)

(b) Calculate the mean, median, mode, variance, standard deviation, skewness and kurtosis of variables monthly salary and balance in saving account. (2 points)

(c) Use box plot to check whether there are outliers among variables loan amount requested, down payment, and EMI. (3 points)

(d) Which variable among continuous variables have high skewness and kurtosis? (3 points)

Question 3 (5 Points)

Local Dhania, the online grocery store, makes a promise that it will deliver the order within 90 minutes. Based on the past data, it is found that the average time taken to deliver is 68 minutes and the corresponding standard deviation is 14 minutes, and follows a normal distribution.

(a) What proportion of orders is delivered after 90 minutes? (1 Points)

(b) If LD would like to ensure that at least 99% of the orders are delivered before 90 minutes, what should be the target mean delivery time (assume that is no change in the standard deviation)? (2 points)

(c) If it is not possible to reduce the mean time to deliver and standard deviation, what should be the promised time to deliver (instead of current 90 minutes) for which the probability of delivery before that time is 99%? (2 points)

Question 4 (5 Points)

The waiting time at the airport security check follows an exponential distribution with a mean of 20 minutes. Ms Thennal K Warrier arrives at the security check 40 minutes prior to the departure of her flight. It takes 5 minutes to reach the boarding gate after clearing the security check and the airline will close the gate 10 minutes prior to the departure of the flight. Passengers not reaching the gate 10 minutes prior to departure are not allowed to board the flight.

(a) What is the probability that Ms Thennal K Warrier will miss the flight? (1 points)

(b) What is the probability that Ms Thennal K Warrier will wait more than 40 minutes at the security check? (1 point)

(c) Ms Thennal K Warrier has been waiting for 20 minute at the security gate. What is the probability that she will wait for another 20 minutes? (1 point)

(d) If Ms Thennal K Warrier would like to ensure that she would not like to miss the flight in 99% cases, how many hours before the flight departure should she reach the airport security? (2 points)

Question 5 (20 points)

Read the case study, "Analytics Empowering Agriculture: Jayalakshmi Agro Tech", and answer the following questions. Use 5% significance where necessary.

1. Anand the cofounder of JAT claims that disease 6 (leaf curl) was accessed at least 30 times every month on average since October 2015. Test at a significance level of 0.05 whether Anand's claim is true. (2 points)
2. It is believed that among the app users for disease information, at least 15% of them access disease information related to disease 6. Test the validity of this belief at 5% significance. (2 points)
3. JAT believes that over the years the average number of app users have increased significantly. Is there a statistical evidence to support that the average number of users during July 2016 and May 2017 is more than number of users during June 2015 and June 2016 at 5% significance. Support your answer with all necessary tests. (3 points)
4. Farmers use Apps to access information about different types of diseases. Using the data, check whether the there is a statistically significant difference in the average number of times various diseases are accessed. (3 points)
5. Anand claims that number of users have increased over a period of two years, he wants to understand If with the number of increasing users, his app usage (number of times his app is accessed in a month by various users) is also increased. Also prove this claim statistically. Also prove statistically if the correlation between users and usage can be non zero. (2 points)
6. A new version of the app was released in the month of Aug 2016, Anand wants to understand after which month in the given time frame post the launch of new version the mean usage pattern starts to show statistically significant shift. (2 points)
7. If a disease is likely to grow in particular weather condition (data given in the disease index sheet), then the access of that disease should be more in the months having suitable weather conditions. Help the analyst in coming up with a statistical test to support the claim for two districts for which the weather and disease access data is given in the data sheet. Mention the diseases for which you can support this claim. Test this claim both for temperature and relative humidity at 90% confidence. (3 points)
8. Based on the data analysis what are the key insights which Anand gets and how they can be used to help farmers. (3 points)

Question 6 (20 Points)

Read the case: ”A Dean’s Dilemma: Selection of Students for the MBA Program”, and answer the following questions.

1. Carryout descriptive analytics (use different data visualization approaches). What insights you are able to gain using descriptive analytics about the students who are placed and not placed? (5 Points)

2. In a random selection of 20 students, what is the probability that exactly 5 students are not placed? What is the probability that at least 5 students are not placed? (5 Points)

3. Consider only the data of students who were placed and answer the following questions: (10 Points)

    a. Use an appropriate test to check whether the student’s marks in SSLC follows a normal distribution. (2 points)

    b. Is there a statistical evidence to suggest that the average salary of students with average score of 60 marks in SSLC is less than the average salary of students with an average score of more than 60 marks in SSLC? Use the appropriate statistical test. (2 points)

    c. The Dean, Easwaran Iyer, believes that the male students earn at least 10000 more than female students per annum. Do an appropriate test to validate this belief. (2 points)

    d. Students from CBSE Board (in SSC) are given higher priority during the admission, is this admission policy justified? Justify your answer. (2 points)

    e. What will be your recommendations to Dr. Easwaran Iyer based on your response to questions 1, 2 and 3. (2 Points)

Question 7 (10 points)

The proportion of undecided voters across 50 districts of a country one-week prior to the election is given in Table 2.

(a) Check whether the undecided votes follow a normal distribution.

(b) Calculate the 90% confidence interval for the proportion of undecided voters.

Table 2 Proportion of undecided voters one week prior to election

12	16	12	10	14	9	8	13	5	5
19	8	6	11	19	14	10	20	11	10
6	6	5	12	16	9	5	9	17	18
15	17	18	13	18	11	7	20	6	11
23	10	24	6	24	18	7	8	5	15

Question 8 (5 points)

An educational psychologist wants to check the claims of some spiritualists that a daily practice of meditation among the students will improve the academic achievement of the students. To control the experiment for academic aptitude, pairs of college students with similar grade point averages (GPA) are randomly assigned to either a group that receives daily training in meditation or a group that doesn't receive training in meditation. At the end of the experiment which lasts for one term, the following GPAs were reported for 10 pairs of participants:

Pair Number	Meditation	No-Meditation
1	4.00	3.75
2	2.65	2.75
3	3.65	3.45
4	2.55	2.11
5	3.20	3.21
6	3.60	3.25
7	2.90	2.58
8	3.41	3.28
9	3.33	3.35
10	2.90	2.65

Use an appropriate hypothesis test to check whether meditation increases academic performance.

Question 9 (10 points)

Hedge funds are alternative investment options that claim to provide better returns compared to investments such as mutual funds and stocks. Hedge funds use many strategies such as *Convertible, Currency, Derivative, Emerging Market, etc.* Siddharth Sinha, an investment advisor at Platinum Investments, strongly believes that the average returns of the hedge fund strategy '*Emerging Market*' is higher than that of '*Derivative*'. A sample of hedge funds that uses these two strategies and their returns are shown in Tables 9.1 and 9.2. Conduct a hypothesis test to check whether the strategy '*Emerging Market*' gives more average returns than the strategy '*Derivatives*'.

Table 9.1 Percentage returns of hedge funds under strategy '*Emerging Market*'

11.20	12.10	13.33	16.40	15.00	10.00	12.00	13.00	12.00	13.00
8.25	7.00	10.00	11.46	11.00	7.70	7.00	12.00	18.00	10.00
13.11	9.00	14.00	9.90	16.00	9.00	6.00	11.40	7.00	16.00
8.41	17.21	14.00	15.00	17.20	18.00	9.00	7.00	15.45	15.00
13.00	18.60	16.00	9.60	12.00	6.00	15.00	8.00	16.29	9.00

Table 9.2 Percentage returns of hedge funds under strategy '*Derivatives*'

17.65	10.20	19.00	14.00	11.00	4.97	11.00	7.00	5.12	4.90
19.00	11.45	16.00	6.87	14.00	8.00	10.78	16.00	18.00	11.00
13.00	17.00	18.00	16.00	12.00	13.26	19.00	10.00	17.00	5.56
8.00	15.55	11.22	6.78	10.00	19.00	14.00	15.00	14.00	7.00
14.00	15.00	18.00	7.78	10.00	15.00	16.20	15.00	11.65	13.00

Question 10 (10 points)

Twenty-five overweight people were assigned three different programs for weight loss, namely: 1. Diet, 2. Exercise and 3. Modification of eating behaviour. The weight changes are recorded after 3 months and are shown in the table below. In the table, positive value indicates weight loss and negative value indicates weight gain.

Table 10.1 Weight loss/Gain Data

S.No	Diet	Exercise	Modification of Eating Behaviour
1	0	-3	10
2	4	-1	1
3	3	8	0
4	5	4	12
5	-3	2	18
6	10	3	4
7	0		-2
8	4		5
9	-2		3
10			4

Use an appropriate hypothesis test to check whether there is statistically significant difference in weight loss between three weight loss programs.

End of the assignment