Self-Assessment (Session 5 & 6)

While answering the questions, think from the perspective of a Data Scientist. If you are in his shoes, how will you respond?

1.For a normal distribution, both the median and mean happen to be 75. Can you guess approximately what will be the value of the mode?

a)150

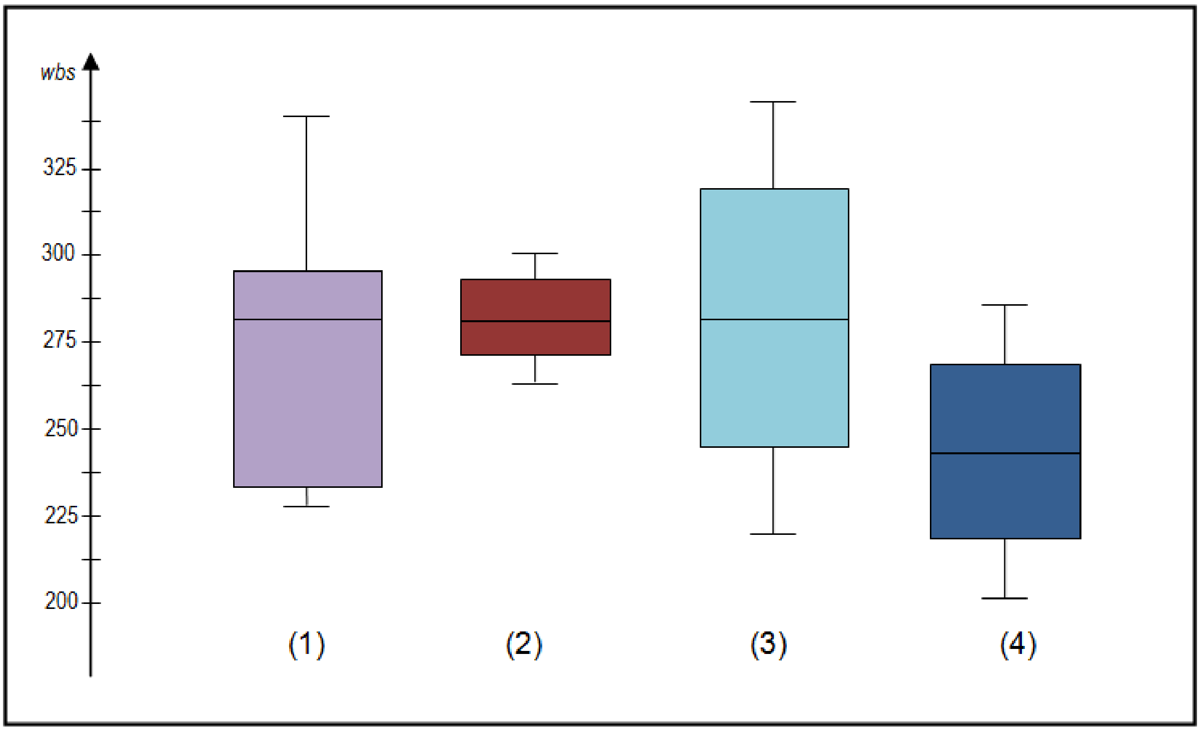
b)0

c)75

d)5625

Solution: In a normal distribution, median, mode and mean are the same. The answer is 75.

2. Which of the below distributions have the highest Inter-Quartile Range?



a)1

b)2

c)3

d)4

Solution: For calculating IQR, we need to consider the upper limit and lower limit of the boxes in the graphs.(3) has the highest IQR of 75 (315 – 240).

3. You are a Data Scientist working with FlipKart. The Category Manager who is tasked with increasing the sales of Sports Merchandise on the site comes up to you and asks

“How are the sales of Sports Merchandise doing this year? Are we better off this year than in the past?”

You have the yearly data of the Sports Merchandise sales for the last decade. You also have monthly data for the six months lapsed this year. This is the only data at your disposal. Which of the below measures will you generate to support your response?

a)Kurtosis
b)Sample Variance
c)Mean/Median/Mode
d)Z-score

Solution: In this scenario, you need to calculate the central tendency numbers as you need to compare apples with apples (historic yearly data with current year data). This is similar to the Patrick example we covered in the class.

4. As a Data Scientist, you are always intrigued by how well you are paid compared with the rest of the folks in other departments within your organization. You somehow got access to an excel that lists out all the monthly salaries of individuals (9500 employees). You quickly calculated the mean and S.D as 65,000/- and 15,000/-. Your salary is 71,000/- What would be your rank in the organization when it comes to salary payout?

a)3325

b)5675

c)4325

d)9500

Solution: In this scenario, you are no longer interested in checking whether you r Salary is better than the average Salary or not. You are looking at the rank of your Salary. So, you need to consider Z-scores.

z-score = (71,000 – 65,000)/ 15000 = 0.4

From the Z-table, you can see that the area that corresponds to 0.4 is 0.6554. This corresponds to the probability that the scores are below 0.4. Multiplying 0.65(rounding) by 9500, we get 6175. This means 6175 salary values are below your Salary value. Subtracting this from 9500, we get 3325 as your rank.

5.Your company is gearing up to organize its annual conference this year. They came to you and asked for an estimate of the approximate number of attendees for this year’s conference so that they can prepare better with the logistics. They gave you the below data. Which number will you quote them as the possible number of attendees for this year’s conference?

Conference Year	**Total number of Attendees**
2008	900
2009	365
2010	900
2011	610
2012	930
2013	340

2014	610
2015	900
2016	714

a)700

b)718

c)900

Solution: In this data set,

Median : 714, Mean : 700, Mode : 900

This is a tricky question as you may be tempted to go with the Mean or for that matter with Median. But you need to pay attention to two things here:

a. Data : You try to observe what is that value that best represents the data set. You see most of the values are hovering in the 900+ range. This is 200 more than the Median and Mean values.
b. Business Case: When you are planning to organize an event, you usually take the 'Better be safe than sorry' approach i.e, you plan extra slots just in case more people turn up.

Taking these two points into consideration, Mode 900 is going to be a better estimate than 700.

6. You are trying to estimate on an average, how many movies does a Data Scientist watch in a year. Since you do not know how many Data Scientists are there in the entire world and the movies-watched data of all these people, you decide to make-do with enquiring your fellow batch mates. You ask 10 of your batch mates the number of movies they watch in a year.

Below is the data:

Student	No.of Movies watched in a year
Ram	12
Shreya	17
Midhun	8
Frank	13
Robert	4
Pulkit	9
Avinash	7
Nidhi	11
Mrinal	15
Nataraj	6

What is the sample variance?

a)15

b)17

c)13

d)19

Solution: Since, this is sample variance, we need to consider 'n-1' in the denominator.

7. A Sales Manager is trying to understand the average size of the deals that his Sales Reps are closing every year. He came up to you and presented the below data. You saw the data and it represented a bell curve. You performed a quick calculation and said "Your sales reps on an average sell somewhere between $282,403 and $419,928." How confident were you in quoting that range?

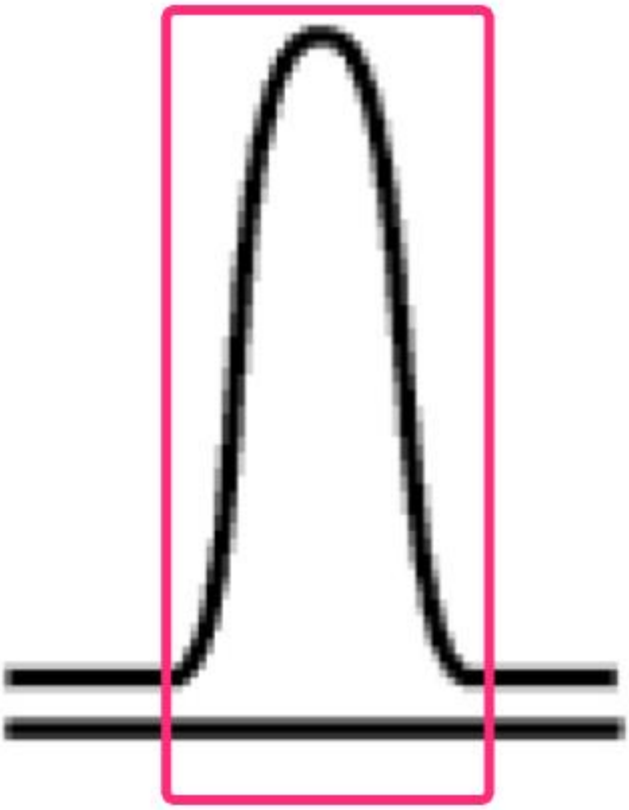
Year	**Total Deals Closed ($)**
2010	359000
2011	321230
2012	290090
2013	343210
2014	383450
2015	361230
2016	399950

a) 75%
b) 68%
c) 95%
d) 99%

Solution: You need to calculate the confidence intervals.

68%	[316784.4	385547]
95%	[282403.1	419928.4]
99%	[248021.7	454309.7]

8.You are working as a Data Scientist at BlueStone.com. When the CEO asks you "Can you mine the data and tell me something interesting?", you get all excited. You collect data about all the purchases made by all the customers on Bluestone.com. You plot a normal distribution curve like the one below and tell the CEO "We need to focus our marketing efforts to the people who made transactions in the highlighted region as they constitute the bulk of the revenue". CEO shakes his head and says, 'Nah..I need to target the people in the regions(tails) outside the highlighted area to grow the business'. Which measure did you overlook and was understood by the CEO?



a)Skewness

b)Kurtosis

c)Sample Variance

d)Mean

Solution:As the CEO of a Jewellery business, you are more interested in acquiring HNI (High Net-worth Individuals) as your customers. For that, you need to analyse the tail of the distribution curve to find those transactions that involved huge sums. Hence, you will be more interested in Kurtosis.

9. You are working on fine-tuning several models to predict the adoption rate of the 'Chat bot' feature in your product. 3 of the models are powered by the Random Forest algorithm, 2 of them are based on Clustering, 5 of them are using Decision tree algorithm. When you tell the Product Manager that you need more time to decide upon a model, he asks you "What are the chances of the Random Forest algorithm successfully predicting the adoption rate?" Your immediate response would be:

a)50%

b) 300%

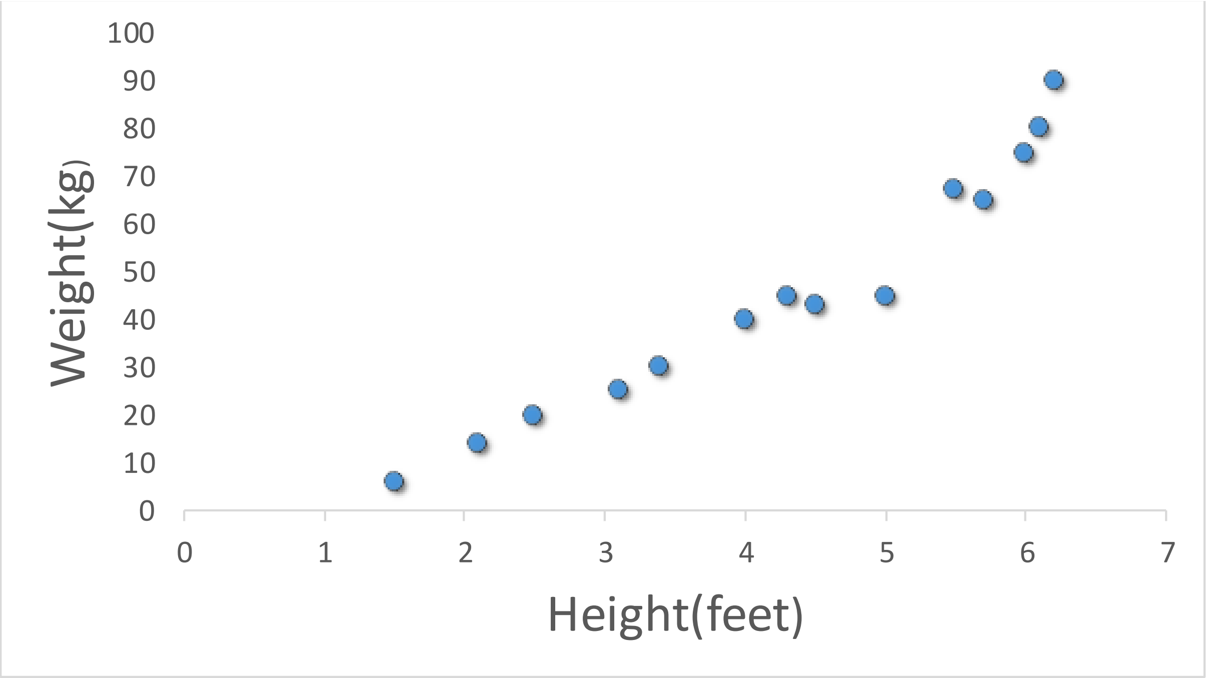
c)40%

d)30%

Solution: As the chances of picking an algorithm are random, you need to calculate the probability.

Probability = (Chances of picking Random Forest Algorithm)/(Chances of picking any algorithm) = (3/10) = 30%

10. From the below graph, what is the approximate average height of people(in feet) who weigh below 40 kgs?



a)2.3

b)2.8

c)3.3

d)2.4

Solution:

Below are the heights of all the bubbles which correspond to weights below 40kgs.

1.5
2.1
2.5
3.1
3.4

Calculating the average, we get it as 2.52. As always is the case with graphs, we select the nearest option to this value, which is 2.4, as answer.