STATISTICS WORKSHEET-1

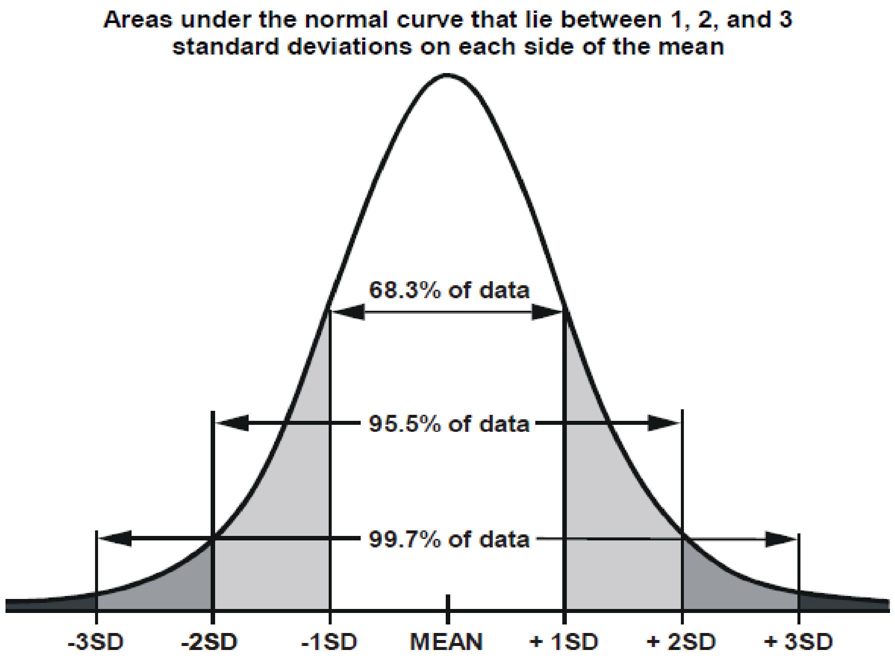
Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. **(A) True**
2. **(A) Central limit Theorem**
3. **(B) Modeling bounded count data**
4. **(D) All of the mentioned**
5. **(C) Poisson**
6. **(B) False**
7. **(B) Hypothesis**
8. **(A) 0**
9. **(C) Outliers cannot conform to then regression relationship**

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10.) A normal distribution is the proper term for a probability bell curve.

- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- In a normal distribution the mean value (average) is also the median (the middle number of a sorted list of data) and the mode (value that appears most often). As this distribution is symmetric about the center, 50% of values are lower than the mean and 50% of values are higher than the mean.
- When we calculate standard deviation we find that generally, 68% of values are within
  1 standard deviation of the mean, 95% of values are within 2 standard deviations of the mean and 99.7% of values are within 3 standard deviations of the mean. The number of standard deviations from the mean is called the z-score, Standard Score, or sigma.


Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean
68.3% of data
95.5% of data
99.7% of data
-3SD
-2SD
-1SD
MEAN
+ 1SD
+ 2SD
+ 3SD

11.) Missing data appear when no value is available in one or more variables of an individual. Due to Missing data, the statistical power of the analysis can reduce, which can impact the validity of the results.

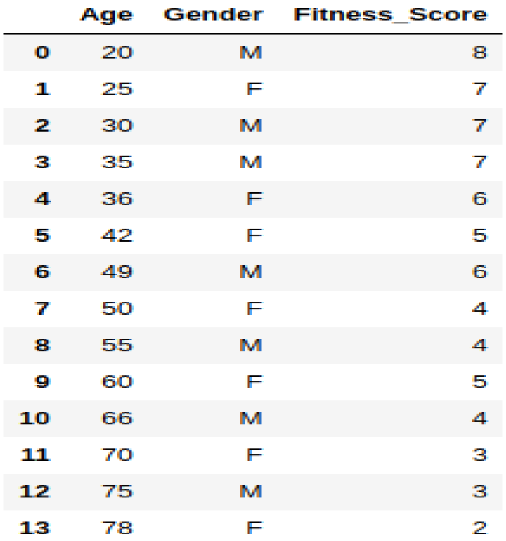
- I used to handle missing data by three imputation techniqes which I also recommended-:
  I. Zero Replacement: Here, you replace the missing value with zero irrespective of everything.
  II. Min or Max Replacement: Replace the missing value with the minimum or maximum value of a feature.
  III. Mean/Median/Mode Replacement: Replace missing value with mean or median or most frequent feature value.

12.) A/B testing also known as split testing is a process of comparing two version of a digital asset to see which one users respond to better. Examples of assets include a landing page, display ad, marketing email, and social post. In an A/B test, half of your audience automatically receives “version A” and half receives “version B.” And you have to analyze that in which version audience interact more. For example in a web page there is a contact button , In A version contact button is on right side and In B version contact button is on left side of web pages for 10 days each and now you have to analyze that in which version your audience interact with contact button more and then take decision that to where put that button.

13.) It is acceptable **when the missing value is not large enough**. But, when the missing values are large enough and you impute them with the mean, the standard errors will be lesser than what they actually would have been.

Means imputation also ignores feature correlation.

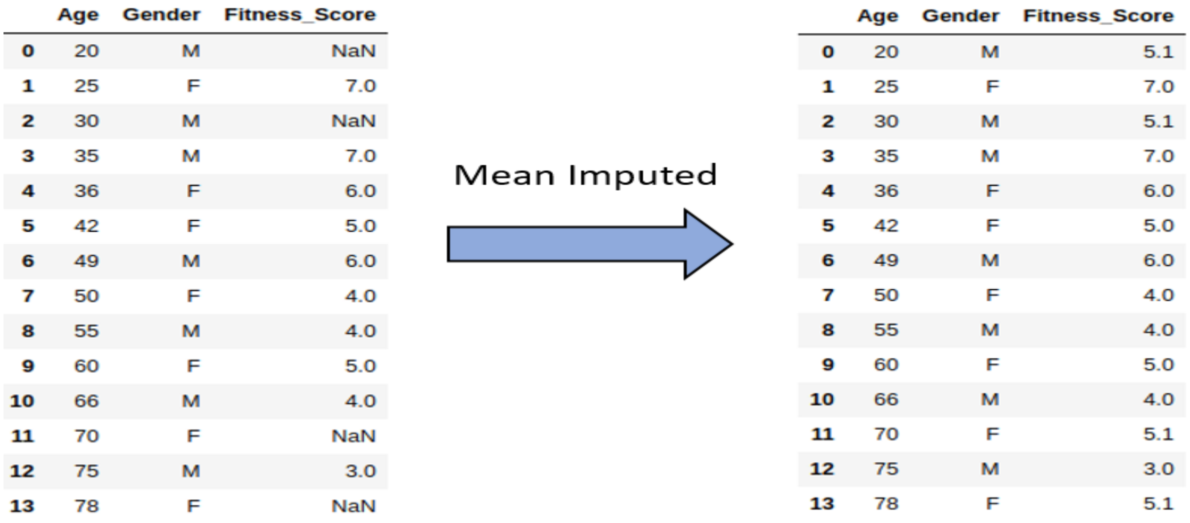
Let’s have a look at a very simple example to visualize the problem. The following table have 3 variables: Age, Gender and Fitness Score. It shows a Fitness Score results (0–10) performed by people of different age and gender.



	Age	Gender	Fitness_Score
0	20	M	8
1	25	F	7
2	30	M	7
3	35	M	7
4	36	F	6
5	42	F	5
6	49	M	6
7	50	F	4
8	55	M	4
9	60	F	5
10	66	M	4
11	70	F	3
12	75	M	3
13	78	F	2

Table with correct, non-missing data

Now let’s assume that some of the data in Fitness Score is actually missing, so that



	Age	Gender	Fitness_Score
0	20	M	NaN
1	25	F	7.0
2	30	M	NaN
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	NaN
12	75	M	3.0
13	78	F	NaN

Mean Imputed

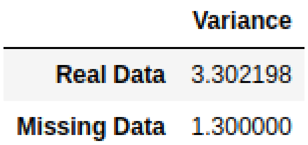
	Age	Gender	Fitness_Score
0	20	M	5.1
1	25	F	7.0
2	30	M	5.1
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	5.1
12	75	M	3.0
13	78	F	5.1

Mean Imputation of the Fitness_Score

Imputed values don’t really make sense — in fact, they can have a negative effect on accuracy when training our ML model. For example, 78 year old women now has a Fitness Score of 5.1, which is typical for people aged between 42 and 60 years old. Mean imputation doesn’t take into account a fact that Fitness Score is correlated to Age and Gender features. It only inserts 5.1, a mean of the Fitness Score, while ignoring potential feature correlations.

Mean reduces a variance of the data

Based on the previous example, variance of the real Fitness Score and of their mean imputed equivalent will differ. Figure below presents the variance of those two cases:



	Variance
Real Data	3.302198
Missing Data	1.300000

Fitness Score variance of the real and mean imputed data

As we can see, the variance was reduced (that big change is because the dataset is very small) after using the Mean Imputation. Going deeper into mathematics, a smaller variance leads to the narrower confidence interval in the probability distribution[3]. This leads to nothing else than introducing a bias to our model.

14.) Linear regression is a linear approach for modelling the relationship between dependent and independent variable. It is commonly used type of predictive analysis. The overall idea of regression is to examine two things:

(1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

(2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

15.) There are two branches of statistics
Descriptive Statistics
The branch of statistics that focuses on collecting, summarizing, and presenting a set of data. For example The mean age of citizens who live in a certain geographical area, the mean length of all books about statistics, the variation in the time that visitors spent visiting a website.

Inferential Statistics
The branch of statistics that analyzes sample data to reach conclusions about a population. For example A survey that sampled 1,264 women found that 45% of those polled considered friends or family as their most trusted shopping advisers and only 7% considered advertising as their most trusted shopping adviser. By using methods discussed in Section 6.4, you can use these statistics to draw conclusions about the population of all women.