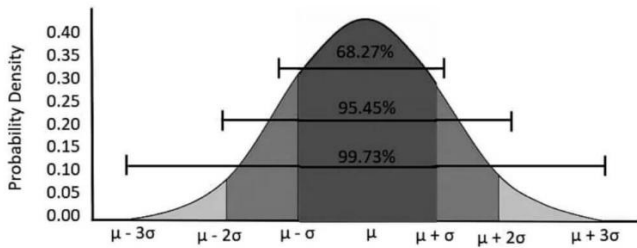


DS_2303		STATISTICS	Worksheet 1
1.	a) True		
2.	a) Central Limit Theorem		
3.	b) Modeling bounded count data		
4.	d) All of the mentioned		
5.	c) Poisson		
6.	b) False		
7.	b) Hypothesis		
8.	a) 0		
9.	c) Outliers can conform to the regression relationship		
10.	<ul style="list-style-type: none"> The Normal Distribution, often known as the Gaussian distribution, is the most important continuous probability distribution in probability theory and statistics. It is also referred to as a bell curve on occasion. In every physical science and in economics, a huge number of random variables are either closely or precisely represented by the normal distribution. Additionally, it can be used to roughly represent various probability distributions, reinforcing the notion that the term "normal" refers to the most common distribution. The normal distribution has the following distinguishing characteristics in all of its various shapes. The bell curves are all symmetric. Skewed distributions are impossible to model using the Gaussian distribution. Mode, median, and the mean are all equivalent. Half of the population falls below the mean, while the other half exceeds it. You may calculate the percentage of data that are within particular ranges of the mean using the Empirical Rule. Therefore, 68% of the values lie within one standard deviation range. 95% of observations lie within two standard deviations, and 99.7% of the values appear within three standard deviations. 		
11.	<ul style="list-style-type: none"> The values or data that are not stored (or not existent) for one or more variables in the provided dataset are referred to as missing data. The dataset's blank spaces display the values that are missing. Missing values in Pandas are typically represented as NaN. Its acronym is Not a Number. Consider carefully each column that has missing values to determine why those values are missing; this information will be necessary to decide how to handle the missing values. 		

	<ul style="list-style-type: none"> • There are 2 main approaches to dealing with missing values: <ol style="list-style-type: none"> 1) Deleting the Missing values 2) Imputing the Missing Values Imputing the Missing Value • There are many imputation methods for replacing the missing values. • You can use different python libraries such as Pandas, and Sci-kit Learn to do this. • Replacing with an arbitrary value • Replacing with the mean • Replacing with the mode • Replacing with the median • Replacing with the previous value – forward fill • Replacing with the next value – backward fill
12.	<ul style="list-style-type: none"> • A/B testing, sometimes referred to as split testing or bucket testing, is a technique for contrasting two iterations of a website or app to see which performs better. • A/B testing, commonly referred to as split testing, is a marketing experiment in which you divide your audience in half to test various campaign iterations and see which one performs best. • In other words, you can show one half of your audience version A of a piece of marketing content while showing the other half version B. • A/B testing is also carried out by experts to obtain insightful information and direct crucial business decisions, such as figuring out which product features are most crucial to customers. A/B testing is a common technique for experimentation in the disciplines of web design and digital marketing.
13.	<ul style="list-style-type: none"> • Mean imputation is typically considered terrible practice since it ignores feature correlation. • Think about the following situation: we have a table containing age and fitness scores, but the fitness score for an eight-year-old is missing. • The elderly person would appear to be far more fit than he actually is if we average the fitness scores of those between the ages of 15 and 80. • Second, mean imputation increases bias while reducing the variance of our data. • The model is less accurate and the confidence interval is smaller as a result of the lower variance. • The variance of the imputed variables is decreased via mean imputation. Since standard errors are reduced by mean imputation, the majority of hypothesis tests and confidence interval calculations are invalidated. • The associations between variables, such as correlations, are not preserved by mean imputation.
14.	<ul style="list-style-type: none"> • A data analysis technique called linear regression uses another related and known data value to estimate the value of unknown data. • It uses a linear equation to quantitatively model the relationship between the unknown or dependent variable and the known or independent variable. • Linear regression is commonly used for predictive analysis and modeling. • For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

	<ul style="list-style-type: none"> The formula for simple linear regression is $Y = mX + b$, where Y is the response (dependent) variable, X is the predictor (independent) variable, m is the estimated slope, and b is the estimated intercept.
15.	<p>Statistics</p> <ul style="list-style-type: none"> Statistics is the branch of mathematics that deals with data. Data (technically a plural word; the singular is 'datum') is a collection of values. A collection of data is often referred to as a data set or set of data, but other words such as a list or simply collection are also often used. Examples of data sets are: Marks in a class test: 9, 2, 5, 8, 10, 3, 5, 8, 8, 9 Inflation rate: 2.1, 3.2, 4.1, 2.3, 5.1, 2.2, 0.5 Voting intention in a referendum: Yes, No, No, Yes, Yes, No. <p>Types of Statistics</p> <ul style="list-style-type: none"> Statistics have majorly categorized into two types: <ul style="list-style-type: none"> (a) Descriptive statistics (b) Inferential statistics <p>(a) Descriptive Statistics</p> <ul style="list-style-type: none"> The data is summarized in this form of statistics using the provided observations. The summary is a representation of a population sample utilizing metrics like the mean or standard deviation. Using tables, graphs, and summary statistics, descriptive statistics is a means to arrange, portray, and describe a collection of data. Consider the number of people utilizing the internet or television in a city. Descriptive statistics are also categorized into four different categories: <ul style="list-style-type: none"> (i) Measure of frequency (ii) Measure of dispersion (iii) Measure of central tendency (iv) Measure of position <p>(b) Inferential Statistics</p> <ul style="list-style-type: none"> Descriptive statistics are interpreted using this type of data. In other words, when the data has been gathered, examined, and summarized, we use these statistics to explain the significance of the data. Or, as another way to put it, it is used to derive inferences from data that is subject to random errors like observational errors, sampling variance, etc. With the help of inferential statistics, we can use data gathered from a sample to extrapolate conclusions about the population. It enables us to make claims that go beyond the scope of the facts or data at hand. Creating estimations, as an illustration, using fictitious research.
