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

1. What do you understand by the term Normal Distribution?

normal distribution The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.

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. Normal distributions are symmetrical, but not all symmetrical distributions are normal. In reality, most pricing distributions are not perfectly normal. ex: a person's weight is normally distributed around a mean of

In [2]:
```
import pandas as pd
data=pd.read_csv(r"C:\Users\Ashwini\Documents\president_heights.csv")
```

In [3]:
```
data
```

Out[3]:

	order	name	height
0	1	George Washington	189
1	2	John Adams	170
2	3	Thomas Jefferson	189
3	4	James Madison	163
4	5	James Monroe	183
5	6	John Quincy Adams	171
6	7	Andrew Jackson	185
7	8	Martin Van Buren	168
8	9	William Henry Harrison	173
9	10	John Tyler	183
10	11	James K. Polk	173
11	12	Zachary Taylor	173
12	13	Millard Fillmore	175
13	14	Franklin Pierce	178
14	15	James Buchanan	183
15	16	Abraham Lincoln	193

	order	name	height
16	17	Andrew Johnson	178
17	18	Ulysses S. Grant	173
18	19	Rutherford B. Hayes	174
19	20	James A. Garfield	183
20	21	Chester A. Arthur	183
21	23	Benjamin Harrison	168
22	25	William McKinley	170
23	26	Theodore Roosevelt	178
24	27	William Howard Taft	182
25	28	Woodrow Wilson	180
26	29	Warren G. Harding	183
27	30	Calvin Coolidge	178
28	31	Herbert Hoover	182
29	32	Franklin D. Roosevelt	188
30	33	Harry S. Truman	175
31	34	Dwight D. Eisenhower	179
32	35	John F. Kennedy	183
33	36	Lyndon B. Johnson	193
34	37	Richard Nixon	182
35	38	Gerald Ford	183
36	39	Jimmy Carter	177
37	40	Ronald Reagan	185
38	41	George H. W. Bush	188
39	42	Bill Clinton	188
40	43	George W. Bush	182
41	44	Barack Obama	185

In [4]:
```
heights=data['height']
```

In [5]:
```
#find, mean,std,min,max
print("mean height         :",heights.mean())
print("standard deviation  :",heights.std())
print("minimum height      :",heights.min())
print("maximum height      :",heights.max())
```

```
mean height         : 179.73809523809524
standard deviation  : 7.015868855358296
```

```
minimum height      : 163
maximum height      : 193
```

1. What is A/B testing?\ Ans--> in the prior section we used statistical inference to make an estimate of a population parameters and measure that the reliability of the eastimate through a confidence interval. in this section, we will explore in detail the use of statistical inference in testing a claim about a population parameter, which is the heart of the scientific method used in research. a/b testing is a way of hypothesis testing

eg : say we want to check the efficacy of new vaccine we create two groups A(control group) - with no vaccine, B(variation group)- with vaccine, after these two groups we check which group is showing better resistance to conclude our hypothesis on new vaccine

In [ ]:

1. What is linear regression in statistics?

If we want to use a variable x to draw conclusions concerning a variable y:y is called dependent or response variable. x is called independent, predictor, or explantory variable. if the relationship between two variables is linear is can be summarrized by a straightline. A straight line can be described by an equation

ex: y=a+bx a is called the intercepr and b the slope of the equation. The slope is the amount by which y increases when x increases by 1 unit.

In [ ]:

1. What are the various branches of statistics?

Ans--> Descriptive Statistics,

```
Inferential Statistics,
```

Ans-->Descriptive statistics is the first part of statistics that deals with the collection of data. People think it is too easy, but it is not that easy. The statisticians need to be aware of the design and experiments. They also need to select the correct focus group and keep away from biases. On the contrary, Descriptive statistics are used to do various kinds of analysis on different studies

ex:The average score of the college students in the math test.The average age of the people who voted for the winning candidate in the last election.The average length of the statistics book.

Descriptive statistics have two parts; -Central tendency measures

mean

mode

midian

-despression of data

range

veriance

standard divation

SKW/precntile

Inferential statistics Inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.

Inference statistics often speak in terms of probability by using descriptive statistics. Besides, a statistician uses these techniques for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are

*Different types of inferential statistics

-Regression analysis

-Analysis of variance (ANOVA)

-Statistical significance (t-test)

-Correlation analysis

In [ ]:

1. How do you handle missing data? What imputation techniques do you recommend?

Ans--> we impute the missing values with the median value when comes to numeric data and also we impute with mode when we have outliers

The problem of missing value is quite common in many real-life datasets. Missing value can bias the results of the machine learning models and/or reduce the accuracy of the model. This article describes what is missing data, how it is represented, and the different reasons for the missing data

*how to hanfdle the missing data. Analyze each column with missing values carefully to understand the reasons behind the missing values as it is crucial to find out the strategy for handling the missing values.

There are 2 primary ways of handling missing values:

*Deleting the Missing values -Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values.

If the missing value is of the type Missing Not At Random (MNAR), then it should not be deleted.

If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted.

The disadvantage of this method is one might end up deleting some useful data from the dataset.

*Imputing the Missing Values here are different ways of replacing the missing values. You can use the python libraries Pandas

In [ ]:

1. Is mean imputation of missing data acceptable practice?

In a Univariate approach, only a single feature is taken into consideration. You can use the class SimpleImputer and replace the missing values with mean, mode, median or some constant value. Multivariate Approach

In a multivariate approach, more than one feature is taken into consideration. There are two ways to impute missing values considering the multivariate approach. Using KNNImputer or IterativeImputer classes.

ex Suppose the feature 'age' is well correlated with the feature 'Fare' such that people with lower fares are also younger and people with higher fares are also older.

In that case, it would make sense to impute low age for low fare values and high age for high fares values. So here we are taking multiple features into account by following a multivariate approach.

Nearest Neighbors Imputations (KNNImputer)

Missing values are imputed using the k-Nearest Neighbors approach where a Euclidean distance is used to find the nearest neighbors. below examples

In [16]:
```
import pandas as pd
df = pd.read_csv('http://bit.ly/kaggletrain', nrows=6)
cols = ['SibSp', 'Fare', 'Age']
X = df[cols]
X
```

Out[16]:

	SibSp	Fare	Age
0	1	7.2500	22.0
1	1	71.2833	38.0
2	0	7.9250	26.0
3	1	53.1000	35.0
4	0	8.0500	35.0
5	0	8.4583	NaN

In [17]:
```
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
impute_it = IterativeImputer()
impute_it.fit_transform(X)
```

Out[17]:
```
array([[ 1.        ,  7.25      , 22.        ],
       [ 1.        , 71.2833    , 38.        ],
       [ 0.        ,  7.925     , 26.        ],
       [ 1.        , 53.1       , 35.        ],
       [ 0.        ,  8.05      , 35.        ],
       [ 0.        ,  8.4583    , 28.50639495]])
```

In [18]:
```
from sklearn.impute import KNNImputer
impute_knn = KNNImputer(n_neighbors=2)
impute_knn.fit_transform(X)
```

Out[18]:
```
array([[ 1.    ,  7.25  , 22.    ],
       [ 1.    , 71.2833, 38.    ],
       [ 0.    ,  7.925 , 26.    ],
       [ 1.    , 53.1   , 35.    ],
       [ 0.    ,  8.05  , 35.    ],
       [ 0.    ,  8.4583, 30.5   ]])
```

In [ ]: