Q1) Identify the Data type for the Following:

| Activity | Data Type |
|---|---|
| Number of beatings from Wife | **Discrete** |
| Results of rolling a dice | **Discrete** |
| Weight of a person | **Continuous** |
| Weight of Gold | **Continuous** |
| Distance between two places | **Continuous** |
| Length of a leaf | **Continuous** |
| Dog's weight | **Continuous** |
| Blue Color | **Discrete** |
| Number of kids | **Discrete** |
| Number of tickets in Indian railways | **Discrete** |
| Number of times married | **Discrete** |
| Gender (Male or Female) | **Discrete** |

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

| Data | Data Type |
|---|---|
| Gender | **Nominal (Discrete)** |
| High School Class Ranking | **Ordinal (Discrete)** |
| Celsius Temperature | **Interval (Continuous)** |
| Weight | **Ratio (Continuous)** |
| Hair Color | **Nominal (Discrete)** |
| Socioeconomic Status | **Ordinal** |
| Fahrenheit Temperature | **Interval (Continuous)** |
| Height | **Ratio (Continuous)** |
| Type of living accommodation | **Nominal (Discrete)** |
| Level of Agreement | **Ordinal (Continuous)** |
| IQ(Intelligence Scale) | **Interval (Continuous)** |
| Sales Figures | **Ratio (Discrete)** |
| Blood Group | **Nominal (Discrete)** |
| Time Of Day | **Interval (Continuous)** |
| Time on a Clock with Hands | **Interval (Continuous)** |
| Number of Children | **Ordinal (Discrete)** |
| Religious Preference | **Nominal (Discrete)** |

| | |
|---|---|
| Barometer Pressure | **Interval (Continuous)** |
| SAT Scores | **Interval** |
| Years of Education | **Ordinal (Continuous)** |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

Ans : **Total outcomes = 2^3 = 8**

**Possible outcomes for two heads and one tail are = (T,T,H), (T,H,T), (H,T,T)**

**Therefore, probability is (3/8)**

Q4) Two Dice are rolled, find the probability that sum is

a) Equal to 1 → **0**
   **Ans: Total number of outcomes = 36**
   **Possible number of outcomes = 0**
   **Therefore, the probability is (0/36) = 0**

b) Less than or equal to 4 → **(1/6)**
   **Ans: Total number of outcomes = 36**
   **Possible number of outcomes = (1,2), (2,1), (1,3), (3,1), (1,1), (2,2)**
   **Therefore, the probability is (6/36) = (1/6)**

c) Sum is divisible by 2 and 3 → **(1/6)**
   **Ans: Total number of outcomes = 36**
   **Possible number of outcomes = (3,3), (6,6), (1,5), (5,1), (2,4), (4,2)**
   **Therefore, the probability is (6/36) = (1/6)**

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

**Ans: Probability that two balls are drawn = $^7C_2$**

**Probability that none of the balls drawn is blue**

$$= \frac{^2C_2 + {}^3C_2 + {}^3C_1 + {}^2C_1}{^7C_2}$$

$$= \mathbf{10/21}$$

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

| CHILD | Candies count | Probability |
|---|---|---|
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

**Ans : Expected value = $\sum x.P(x) = (1*0.015) + (4*0.20) + (3*0.65) + (5*0.005) + (6*0.01) + (2*0.120)$**

**$= 3.09$**

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

- For Points,Score,Weigh>
  Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Ans:** **Python code :**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_csv('./Downloads/Q7.csv')

Points_mean = data['Points'].mean()
Points_std = data['Points'].std()
Points_median = data['Points'].median()

Score_mean = data['Score'].mean()
Score_std = data['Score'].std()
Score_median = data['Score'].median()

Weigh_mean = data['Weigh'].mean()
Weigh_std = data['Weigh'].std()
Weigh_median = data['Weigh'].median()

print('Points_mean : ',Points_mean)
print('Points_std : ',Points_std)
print('Points_median : ',Points_median)

print('===============')

print('Score_mean : ',Score_mean)
print('Score_std : ',Score_std)
print('Score_median : ',Score_median)

print('===============')

print('Weigh_mean : ',Weigh_mean)
print('Weigh_std : ',Weigh_std)
print('Weigh_median : ',Weigh_median)
```

| - | **Points** | **Score** | **Weight** |
|---|---|---|---|
| Mean | 3.596 | 3.217 | 17.848 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.891 | 3.54 | 17.43 |
| Variance | 0.285 | 0.957 | 3.19 |
| Standard Deviation | 0.534 | 0.978 | 1.786 |
| Range | [2.76 , 4.93] | [1.513 , 5.424] | [14.5 , 22.9] |

**Use Q7.csv file**

Q8) Calculate Expected Value for the problem below

a) The weights (X) of patients at a clinic (in pounds), are
108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

**Ans: = (180+110+123+134+135+145+167+187+199)/9**

**= 145.333**

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Ans : Python Code:**

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import skew, kurtosis

df = pd.read_csv('./Downloads/Q9_a.csv')
array = df.values

print(skew(array, axis = 0, bias = True))
print(kurtosis(array, axis = 0, bias = True))

del df['Index']

sns.set_style('darkgrid')
plt.boxplot(df);
```

**Cars speed :**

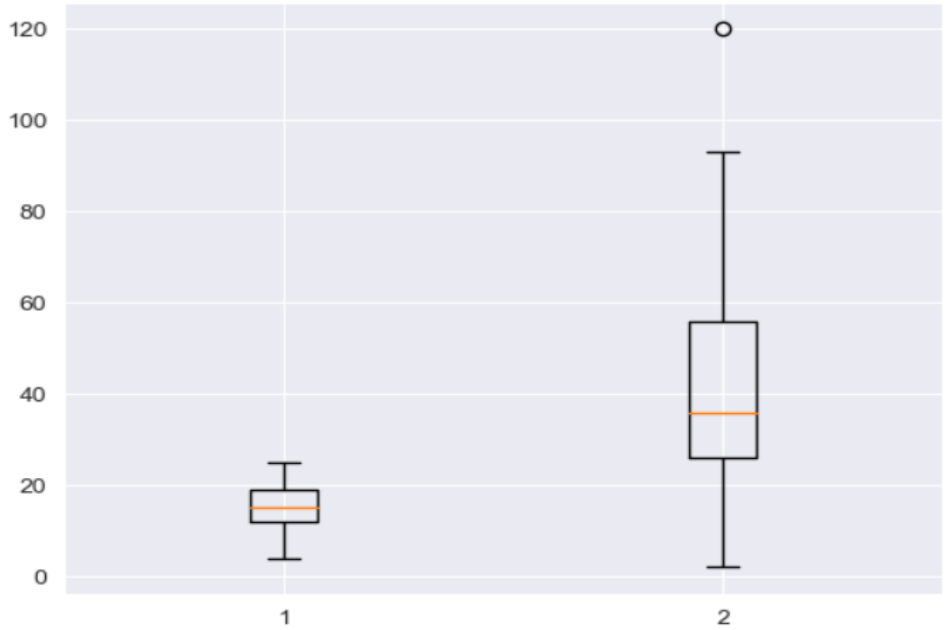
Skewness-> -0.113955

Kurtosis -> 0.81

**distance :**

Skewness-> 0.81

Kurtosis -> 0.41



**Use Q9_a.csv**

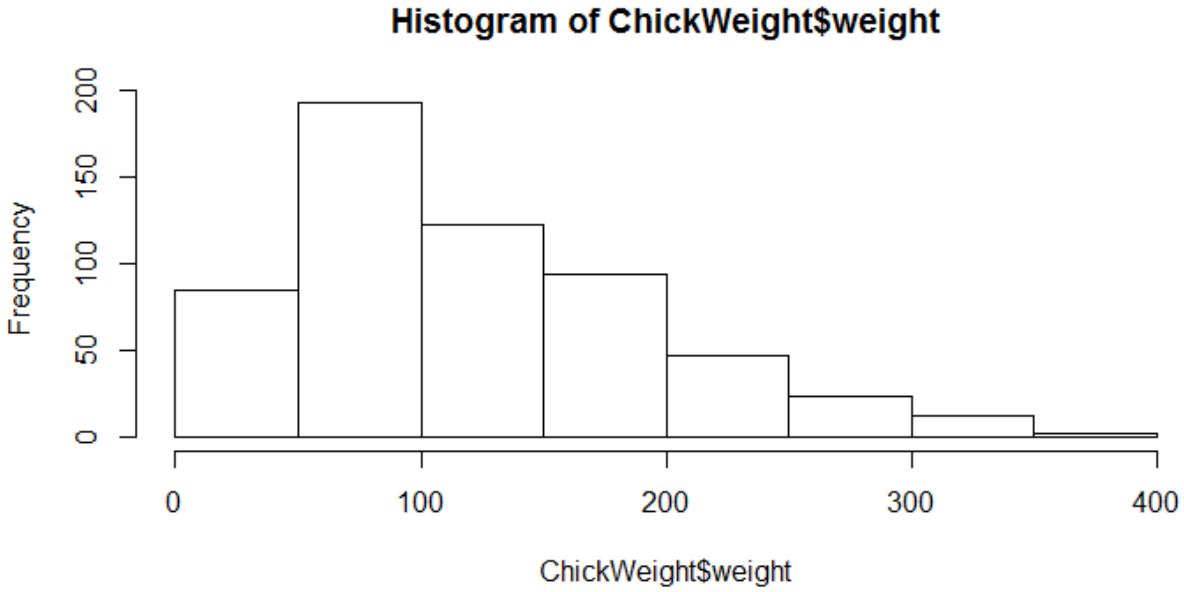
**SP and Weight(WT)**

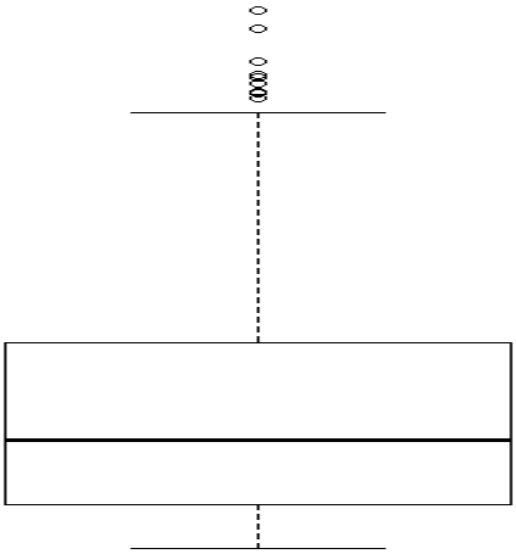
| **Ans :** | **SP :** | **Weight :** |
|---|---|---|
| | Skewness-> 1.61 | Skewness-> 1.61 |
| | Kurtosis -> 0.95 | Kurtosis -> 0.95 |

**Use Q9_b.csv**

**Q10) Draw inferences about the following boxplot & histogram**

---





**Ans: In histogram graph, peak is right skewed and tail is on right side and outliers present on maximum side.**

**In boxplot graph, outliers present on the maximum side.**

**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

**Ans: Python code:**

```
from scipy import stats

print(stats.norm.ppf(0.94))
print(stats.norm.interval(0.94, 200, 30/pow(2000,0.5)))

print(stats.norm.ppf(0.96))
print(stats.norm.interval(0.96, 200, 30/pow(2000,0.5)))

print(stats.norm.ppf(0.98))
print(stats.norm.interval(0.98, 200, 30/pow(2000,0.5)))
```

| CI | Range |
|---|---|
| 94% | [198.74 , 201.26] |
| 96% | [198.62 , 201.38] |
| 98% | [198.43 , 201.56] |

**Q12)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1) Find mean, median, variance, standard deviation.

**Ans : Python code:**

```
import numpy as np
a = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]
print(np.mean(a))
print(np.median(a))
print(np.std(a))
print(np.var(a))
```

**mean = 41**

**Mode = 40.5**

**Variance = 24.11111**

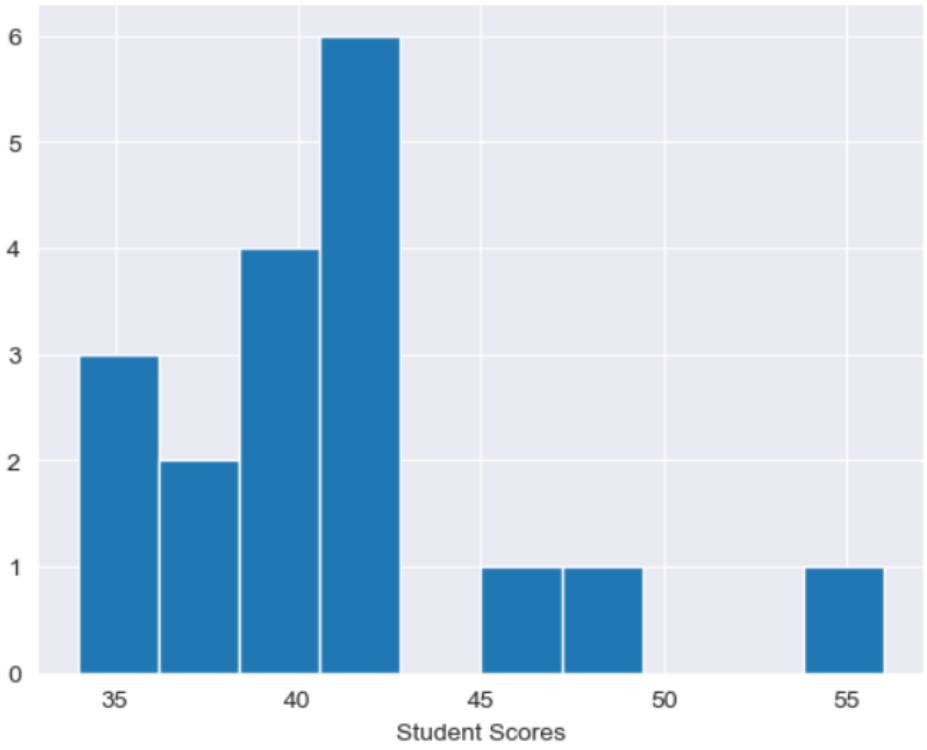
**SD = 4.9103**

2) What can we say about the student marks?

**Ans : Data is slightly right skewed.**

**Python code:**

```
a = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]
plt.hist(a)
plt.xlabel('Student Scores')
```




Q13) What is the nature of skewness when mean, median of data are equal?

**Ans: When mean and median of the data are equal, then the data set is perfectly normal. So, skewness is zero.**

Q14) What is the nature of skewness when mean > median ?

**Ans: Negative skewness (more weight in right tail of the distribution)**

Q15) What is the nature of skewness when median > mean?

**Ans: Positive skewness (more weight in left tail of the distribution)**

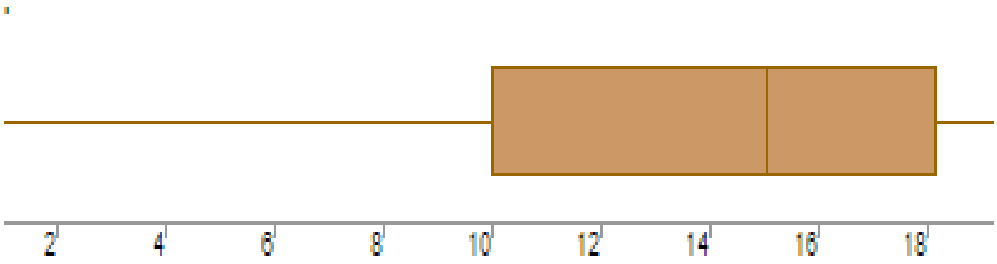
Q16) What does positive kurtosis value indicates for a data ?

**Ans: Distribution is peaked and represents thicker tails**

Q17) What does negative kurtosis value indicates for a data?

**Ans: Wider peaked and represents thinner tails**

Q18) Answer the below questions using the below boxplot visualization.




What can we say about the distribution of the data?

**Ans: Data is not distributed normally as the median is present on higher value**

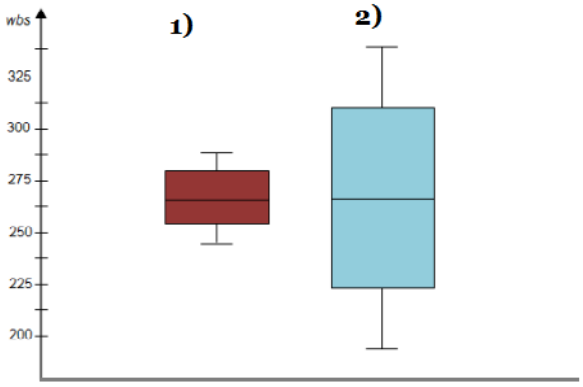
What is nature of skewness of the data?

**Ans: Negative skewness. Data is skewed towards left.**

What will be the IQR of the data (approximately)?

**Ans: Q3 – Q1 = 10-18 = 8**

Q19) Comment on the below Boxplot visualizations?




Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

**Ans: Data is like normally distributed and the median is same in both boxplots.**

**No outliers are present in the Boxplot 1. No skewness is present.**

Q 20) Calculate probability from the given dataset for the below cases
Data _set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

a. P(MPG>38)
b. P(MPG<40)
c. P (20<MPG<50)

**Ans: Python code:**

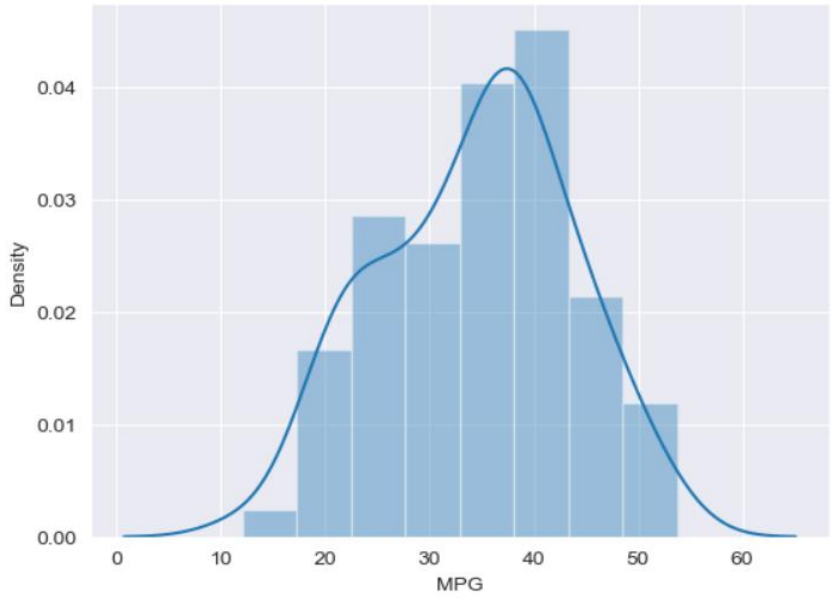
```
cars = pd.read_csv('./Documents/Datasets/Cars.csv')
mean = cars['MPG'].mean()
std = cars['MPG'].std()

# P(MPG > 38)
print(1 - stats.norm.cdf(38, mean, std))

# P(MPG < 40)
print(stats.norm.cdf(40, mean, std))

# P(20<MPG<50)
print(stats.norm.cdf(50,mean,std) - stats.norm.cdf(20,mean,std))

sns.distplot(cars['MPG'], kde = True)
```

P(MPG>38) → **0.34759018**
P(MPG<40) → **0.7293527**
P (20<MPG<50) → **0.013**

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv

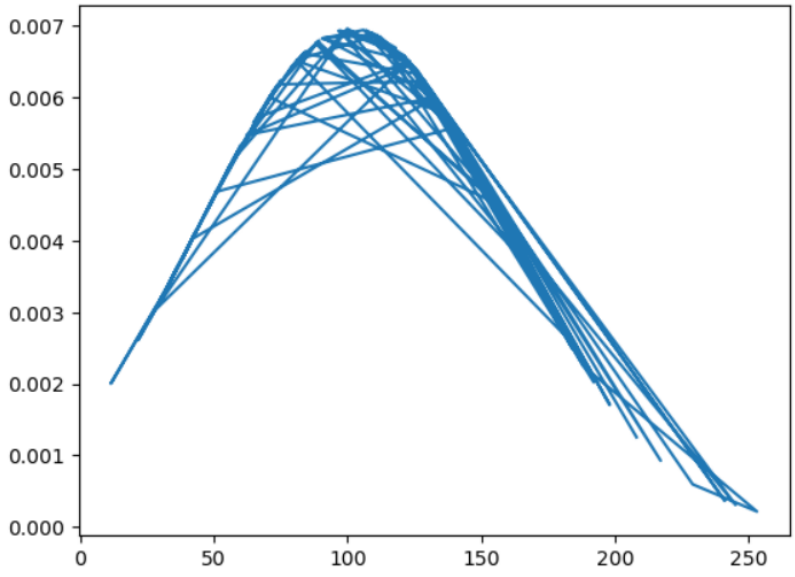
**Ans: (Normally distributed)**




b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution
Dataset: wc-at.csv

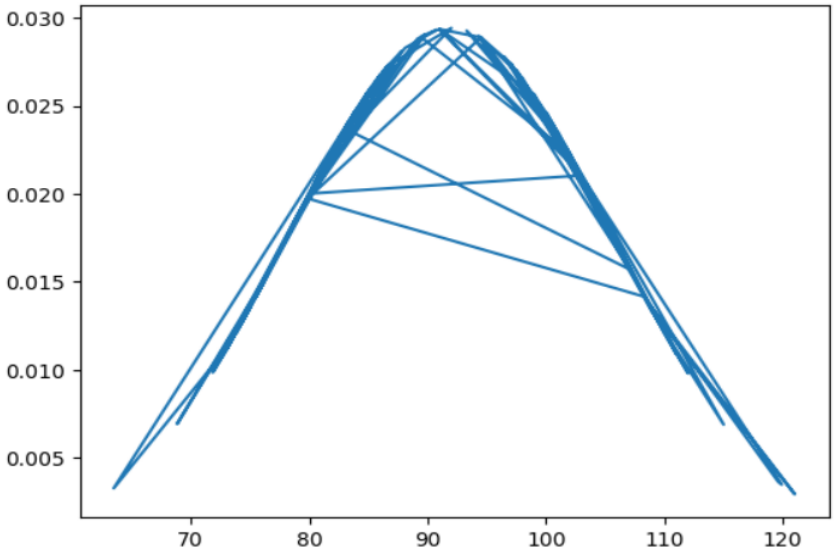
**Ans: 'AT' → Distributed normally**

```
import pandas as pd
cars = pd.read_csv('./Documents/Datasets/WC_AT.csv')

mean_AT = cars['AT'].mean()
std_AT = cars['AT'].std()

from scipy import stats
plt.plot(cars['AT'], stats.norm.pdf(cars['AT'], mean_AT, std_AT))
```




**'Waist' → Normally distributed**

```
import pandas as pd
cars = pd.read_csv('./Documents/Datasets/WC_AT.csv')

mean_waist = cars['Waist'].mean()
std_waist = cars['Waist'].std()

from scipy import stats
plt.plot(cars['Waist'], stats.norm.pdf(cars['Waist'], mean_waist, std_waist))
```



Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

**Ans: Code:**

```
# Z-Scores
# For 60% CI
stats.norm.ppf(0.2)

# For 90% CI
stats.norm.ppf(0.05)

# For 90% CI
stats.norm.ppf(0.03)
```

| Confidence interval | Z scores |
|---|---|
| 60% | -0.8416212 |
| 90% | -1.644854 |
| 94% | -1.880794 |

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

**Ans: t-test is used as sample size is less than 25**

**Code:** stats.t.ppf (x, 24)

| Confidence interval | T scores |
|---|---|
| 95% | -2.063899 |
| 96% | -2.171545 |
| 99% | -2.79694 |

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode → pt(tscore,df)

df → degrees of freedom

**Ans: Python code:**

```
t=(260-270)/(90/18**0.5)
p_value=1-stats.t.cdf(abs(t),df=17)
p_value
```

**0.3218**