|  |  |
| --- | --- |
| 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 |

Q1) Identify the Data type for the Following:

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

Nominal, Ordinal, Interval, Ratio.

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

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

## ANS – Thus, the probability of getting two heads and one tail on tossing three coins at once is equal to 3/8.

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

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

## Sol- a) Zero

## b) 1/6

## c)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?

## Sol – 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

## Sol – Expected no.of candies for a randomaly selected child =

## 1\*0.015 + 4\*0.2 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.12

## = 3.09

## Therefore, expected no.of candies for the randomly selected child is 3 candies.

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

* For Points,Score,Weigh>

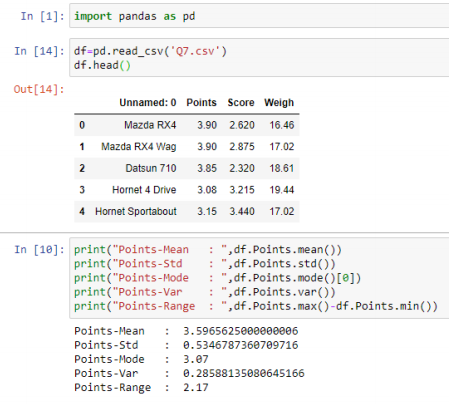
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

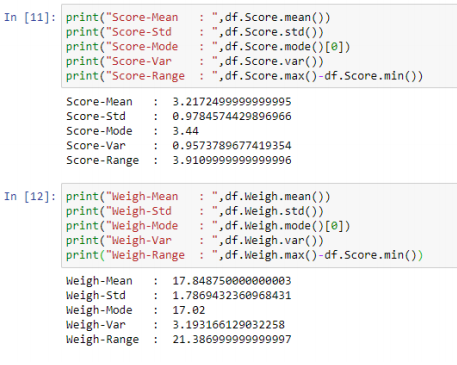
**Use Q7.csv file**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weight** |
| **Mean** | 3.596 | 3.217 | 17.84 |
| **Median** | 3.695 | 3.325 | 17.710 |
| **Median** | 3.07& 3.92 | 3.44 | 17.02 & 18.90 |
| **Std** | .5346 | .9784 | 1.786 |
| **Vairance** | .2858 | .9573 | 3.19 |
| **Range** | 2.17 | 3.910 | 8.399 |

**Inference : 1.Mean & Median approx equal therefore data is normally distributed**

**2. Variance is less , all data points are equal to each other.**





Q8) Calculate Expected Value for the problem below

1. 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?

## Sol:

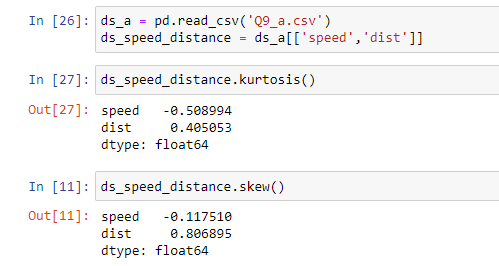
## Expected Value = 145.33(Its nothing but avg of the given data)

## (108+110+123+134+135+145+167+187+199)/9=1308/9= 145.333

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

**Cars speed and distance**

**Use Q9\_a.csv**



SP and Weight(WT)

Inferences from skewness

Negative skewness ( -0.117510) of speed indicates that data is left skewed. Ie median>mean. Ie more values of speed are greater than the average speed and extreme values of speed are lesser than the mean.

Since, the skewness ( -0.117510)  is between (-0.5) to (0.5). Therefore data of speed is fairly symmetrical.

Positive skewness (0.806895) of dist indicates that it is right skewed that is  mean > median. Ie. More values of dist are lower than the average dist and extreme values of dist are greater than the mean.

Since, the skewness (0.806895) is between 0.5 & 1, therefore data of dist column is moderately skewed.

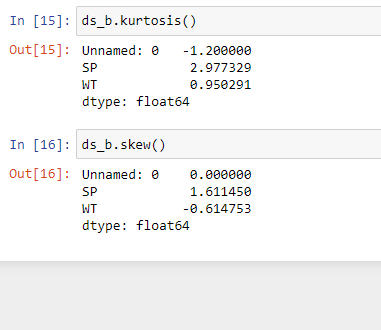
Inferences from kurtosis

Negative kurtosis (- 0.508994 ) indicates light-tailed distribution. Ie. The speed values has no outliers.

Positive kurtosis ( 0.405053 ) indicates heavy-tailed distribution .ie. the dist  values  has large outliers.

SP and Weight(WT)

**Use Q9\_b.csv**



Inferences from skewness

Negative skewness ( -0.117510) of speed indicates that data is left skewed. Ie median>mean. Ie more values of speed are greater than the average speed and extreme values of speed are lesser than the mean.

Since, the skewness ( -0.117510)  is between (-0.5) to (0.5). Therefore data of speed is fairly symmetrical.

Positive skewness (0.806895) of dist indicates that it is right skewed that is  mean > median. Ie. More values of dist are lower than the average dist and extreme values of dist are greater than the mean.

Since, the skewness (0.806895) is between 0.5 & 1, therefore data of dist column is moderately skewed.

Inferences from kurtosis

Negative kurtosis (- 0.508994 ) indicates light-tailed distribution. Ie. The speed values has no outliers.

Positive kurtosis ( 0.405053 ) indicates heavy-tailed distribution .ie. the dist  values  has large outliers.

SP and Weight(WT)

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



## Ans: 1.Data distribution is +ve skew.

## 2. Most of data lies in btn 50-100.



Ans: 1. Outliers prenent on uooerextreme.

2. Thete are 7 outliers on upper extreme .

3. Data is +ve skew.

**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?

## Average @ 98 : 198.439 Sol: stats.norm.interval(CI,sample mean,SD/sqrt(n))

## Sample mean = 2000

## SD = 30

## N = 2000

## Average @ 94 : 198.73

## Average @ 96 : 198.622

**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: mean = 41.0 median = 40.5 variance = 24.11 std = 4.91

1. What can we say about the student marks?

## Ans: Since variance is more , score of student are not equal to each other

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

## Ans: Skewness will be zero.

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

## Ans: Data distribution is positively skewed.

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

## Ans: Data distribution is nigitively skewed.

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

## Ans: When compared to Normal distribution graph, in positive kurtosis the distribution of data will be heavier at tails and sharp peak will be there.

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

## Ans: When compared to Normal Distribution graph, in positive kurtosis the distribution of data will be flat tails will be there.

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



What can we say about the distribution of the data?

## 1. Distribution of the data is towards the tail

## 2. 1st 25% data lies between = 2- 10

## 3. 50% data lies between = 10-18

## 4. 2nd 25% data lies between = 18 –

What is nature of skewness of the data?

## 1. Left whisker length is more therefore data is leftskewed

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

## Sol: Q1=10

## Q2 = 14.6 approx

## 18

## Q1=18-10=8 Approx

Q19) Comment on the below Boxplot visualizations?



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

## Ans: 1. Both have similar median

## 2. Both are normally distributed

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

* 1. P(MPG>38)

## 1-stats.norm.cdf(38,df.MPG.mean(),df.MPG.std()) =.3475=34.75%

* 1. P(MPG<40)

## stats.norm.cdf(40,df.MPG.mean(),df.MPG.std()) = .7293=72.93%

c. stats.norm.cdf(50,df.MPG.mean(),df.MPG.std()) -

## stats.norm.cdf(20,df.MPG.mean(),df.MPG.std()) = .8988=89.88% P

## (20<MPG<50)

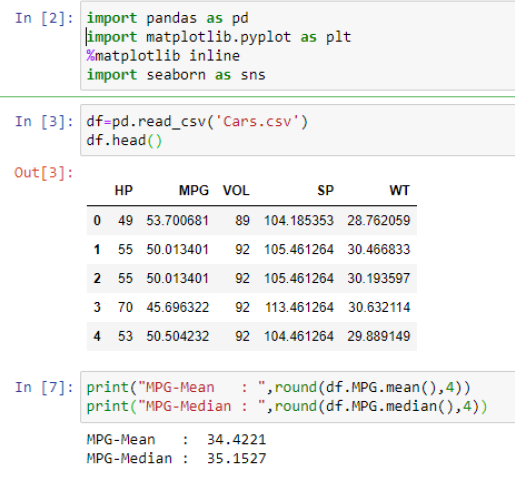
## 

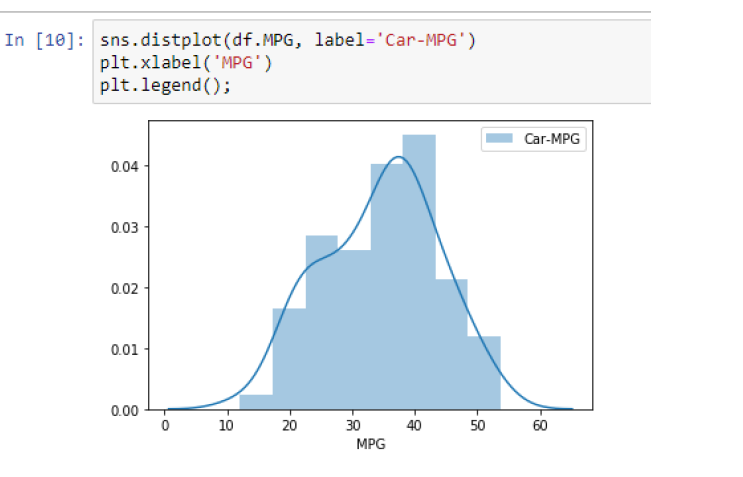


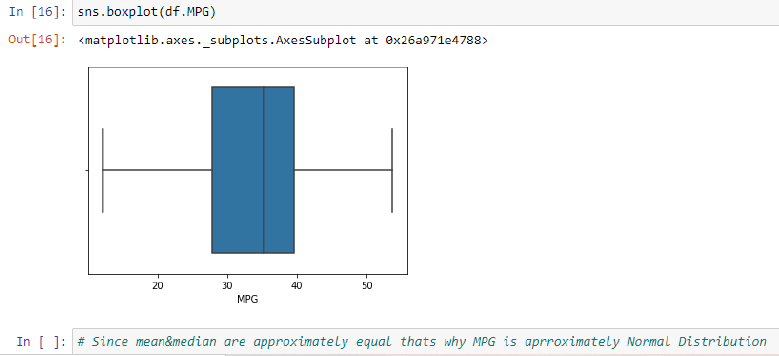
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

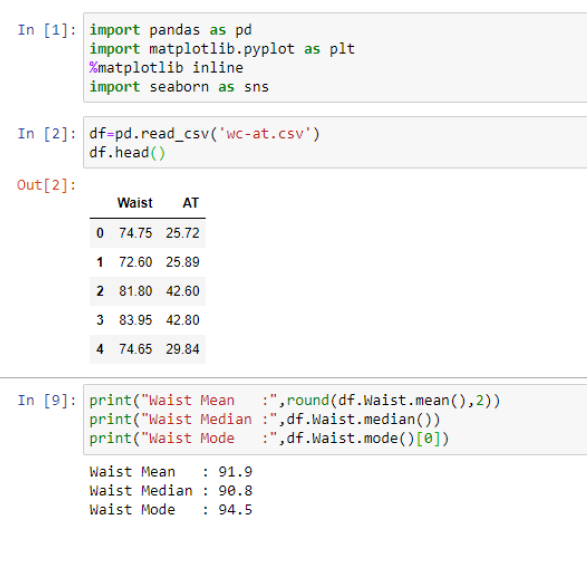


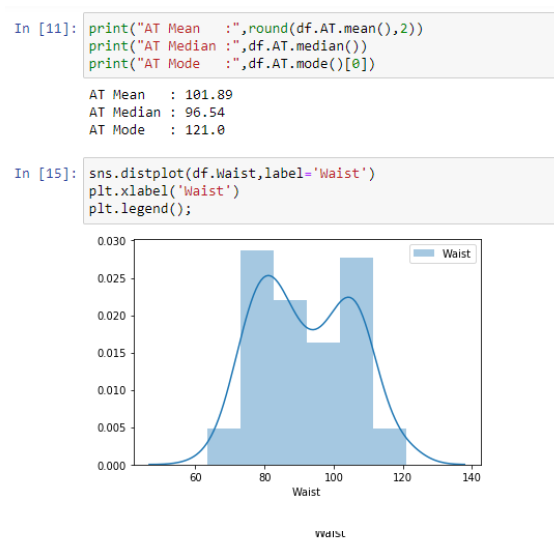


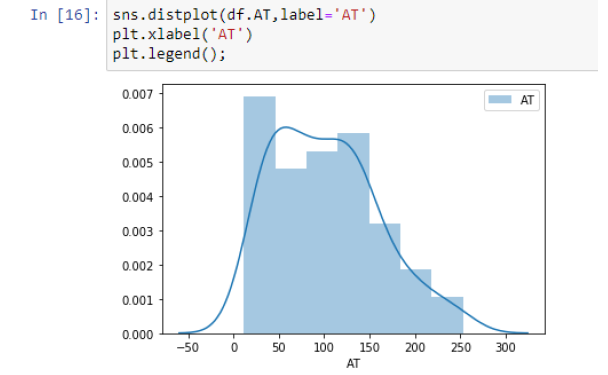


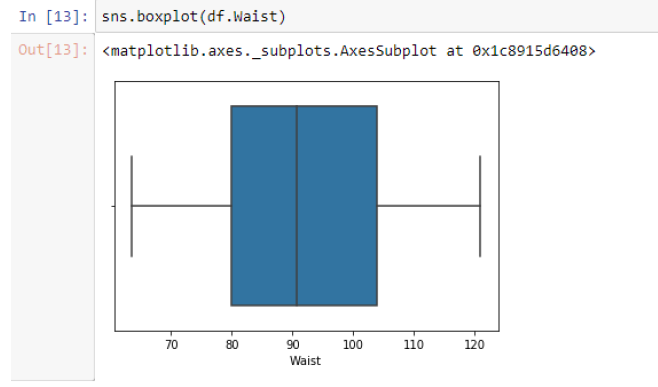
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

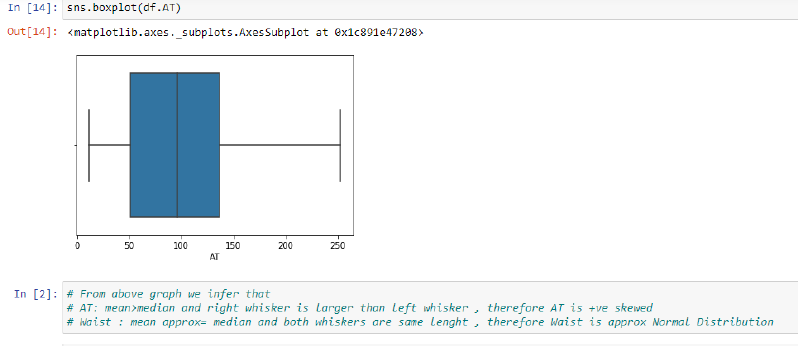
Dataset: wc-at.csv











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

## Ans: @90%

## (1-α)=90%=.9 α=.1

## since we consider α/2= .05 (1-.05)= .95 therefore Z=1.65 from z table @ & (1-α/2)

## Z @ 94% = 1.89

## Z @ 60% = 0.85

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

## Ans : stat.t.ppf(CI,SampleSize) - Python

## t @ 95% = 1.708 t @ 96% = 1.824 t @ 99% = 2.485

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 : 0.3216 by using code stats.t.cdf(t,n) t = -.471

## t score is cal using t score formula

## X= mean of sample =260 Mue = total ppln = 270 n= sample = 18 s = sd=90

