|  |  |
| --- | --- |
| 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 | qualitative |
| Number of kids | discrete |
| Number of tickets in Indian railways | discrete |
| Number of times married | discrete |
| Gender (Male or Female) | Qualitative |

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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | NOMINAL |
| Time Of Day | Interval |
| Time on a Clock with Hands | Ratio |
| Number of Children | ORDINAL |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained? 3/8

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

1. Equal to 1 Zero
2. Less than or equal to 4 3/12=1/6
3. Sum is divisible by 2 and 3 5/36

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?

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

Expected number of candies for a randomly selected child

=  1 \* 0.015  + 4\*0.20  + 3 \*0.65  + 5\*0.005  + 6 \*0.01  + 2 \* 0.12

= 0.015 + 0.8  + 1.95 + 0.025 + 0.06 + 0.24

=       3.090

=  3.09

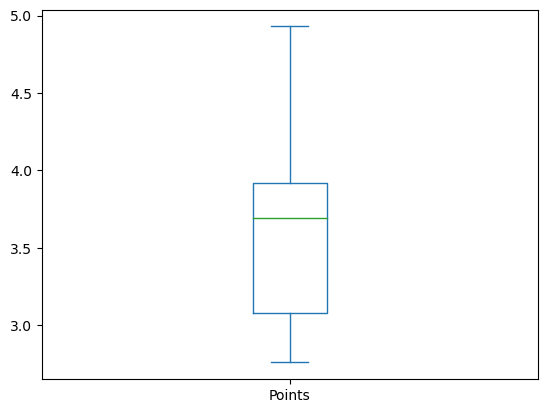
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

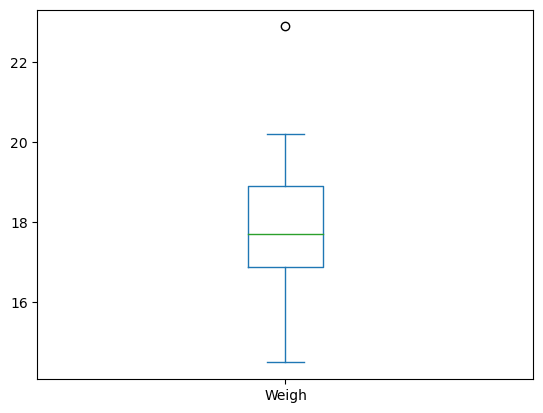
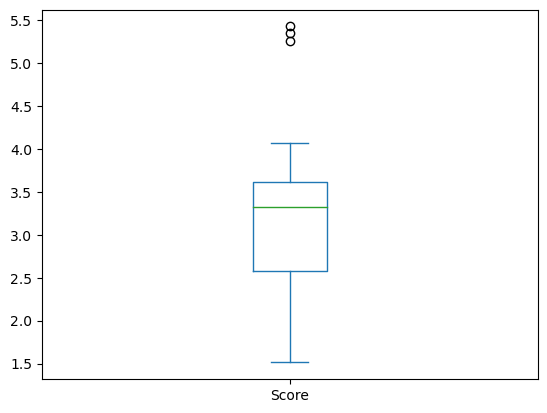
* For Points,Score,Weigh>

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

**Use Q7.csv file**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Mode** | **Variance** | **Std** | **Range** |
| **points** | **3.59** | **3.69** | **3.07,3.92** | **0.28** | **0.53** | **2.17** |
| **score** | **3.21** | **3.32** | **3.44** | **0.95** | **0.97** | **3.91** |
| **weigh** | **17.84** | **17.710** | **17.02,18.90** | **3.19** | **1.78** | **8.4** |

****

****

A **standard deviation**  is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out. A standard deviation close to zero indicates that data points are close to the mean, whereas a high or low standard deviation indicates data points are respectively above or below the mean.

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?

Expected Value  =  ∑ ( probability  \* Value )

  ∑ P(x).E(x)

there are 9 patients

Probability of selecting each patient = 1/9

Ex  108, 110, 123, 134, 135, 145, 167, 187, 199

P(x)  1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  1/9

Expected Value  =  (1/9)(108) + (1/9)110  + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199

= (1/9) ( 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)

= (1/9)  (  1308)

= 145.33

Expected Value of the Weight of that patient = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SKEW**

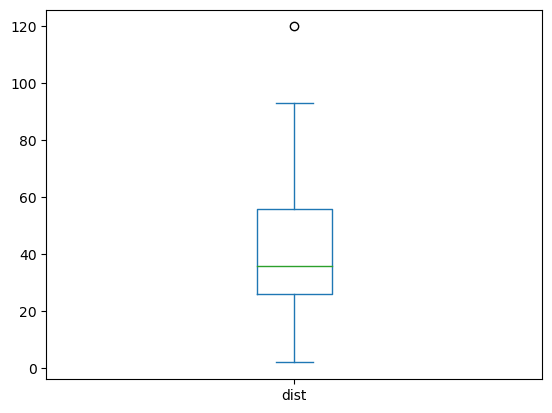
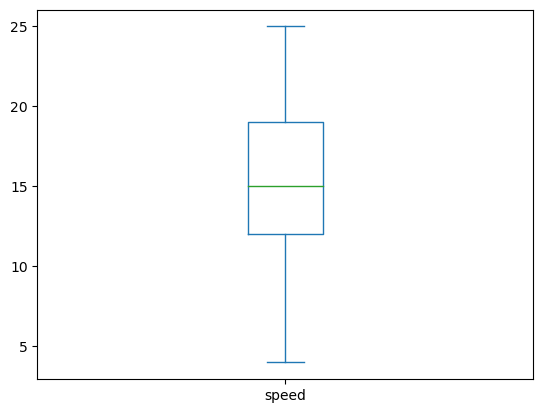
speed -0.117510

dist 0.806895

**KURT**

speed -0.508994

dist 0.405053

****

**SP and Weight(WT)**

**Use Q9\_b.csv.**

**skew**

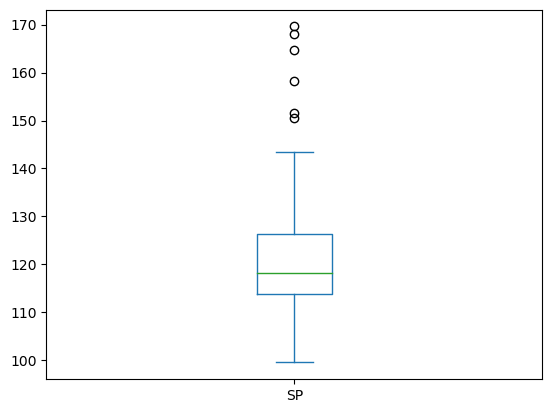
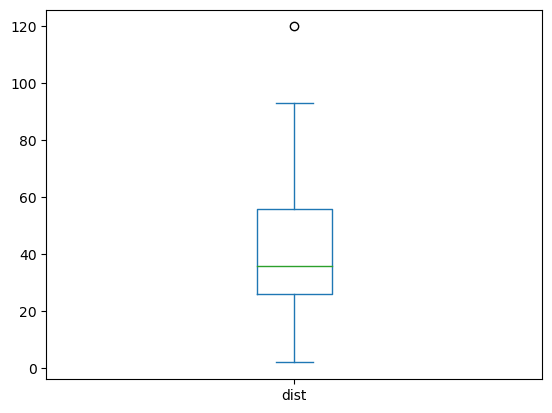
SP 1.611450

WT -0.614753

**KURT**

SP 2.977329

WT 0.950291

****

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



It is a right skewed data and having extreeme values

**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?

X+/-(Z1- α. σ/sqrt(n)

Degrees of freedom= 2000-1= 1999

Confidence interval= 94% (1- σ/2)= 1-0.03) =0.97

for confidene interval for 94% is 1.882

Confidence interval for 98%= 2.33

Confidence interval for 96% = 2.05

**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.

Mean 41 ,median 40.5,variance 25.52,standard deviation 5.052.

1. What can we say about the student marks?

Mean and median are at near same point It is symmetrical data

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

When the mean and median of a dataset are equal, it is an indication that the dataset is symmetrical, or "normally distributed". In this case, there is no skewness.

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

When the mean is greater than the median, the distribution is said to be positively skewed. This means the data is concentrated on the lower end and the tail of the distribution is longer on the right side.

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

When the median is greater than the mean, the distribution is said to be negatively skewed. This means the data is concentrated on the higher end and the tail of the distribution is longer on the left side.

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

A positive kurtosis value indicates that the data has a heavy tail, meaning that there are more values at the extremes of the distribution than would be expected from a normal distribution. This could signify that the data is more prone to outliers or extreme values.

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

A negative kurtosis value indicates that the data has a light tail, meaning that there are fewer values at the extremes of the distribution than would be expected from a normal distribution. This could signify that the data is less prone to outliers or extreme values.

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



What can we say about the distribution of the data?

The above Boxplot is not normally distributed the median is towards the higher value

What is nature of skewness of the data?

The data is a skewed towards left. The whisker range of minimum value is greater than maximum

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

IQR = Q3-Q1  
18-10=8  
  
Q19) Comment on the below Boxplot visualizations?



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

First there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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)= 0.3475939251582705

b.P(MPG<40)= 0.7293498762151616

c.P (20<MPG<50)= 1.2430968797327613e-05

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Normally distributed

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

Dataset: wc-at.csv

follows

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

*Z-score for 90% confidence interval*

stats.norm.ppf(0.9)

Out[6]:

1.2815515655446004

In [7]:

*## Z-score for 94% confidence interval*

stats.norm.ppf(0.94)

Out[7]:

1.5547735945968535

*# Z-score for 60% confidence interval*

stats.norm.ppf(0.6)

Out[8]:

0.2533471031357997

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

**from** scipy **import** stats

**from** scipy.stats **import** norm

In [5]:

*# t scores of 95% confidence interval for sample size of 25*

stats**.**t**.**ppf(0.975,24) *# df = n-1 = 24*

Out[5]:

2.0638985616280205

In [9]:

*# t scores of 96% confidence interval for sample size of 25*

stats**.**t**.**ppf(0.98,24)

Out[9]:

2.1715446760080677

In [11]:

*# t scores of 99% confidence interval for sample size of 25*

stats**.**t**.**ppf(0.995,24)

Out[11]:

2.796939504772804

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

from scipy import stats

from scipy.stats import norm

# Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

# Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days

# find t-scores at x=260; t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))

t= (260-270)/(90/18\*\*0.5)

t

-0.4714045207910317

# Find P(X>=260) for null hypothesis

# p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... Using cdf function

p\_value = 1-stats.t.cdf(abs(-0.471),df = 17)

p\_value

0.32181403316850754

# OR p\_value=stats.t.sf(abs(t\_score),df=n-1)... Using sf function

p\_value = stats.t.sf(abs(-0.471),df = 17)

p\_value

0.3218140331685075