| 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 | Ordinal |
| 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: probability = no of instances for 2 heads & 1 tail / total no of instances = 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

Ans: a.) Probability = Expected instances/total instances = 0/36 = 0

b.) Probability = Expected instances/total instances = 6/36 = 1/6

c.) Probability = Expected instances/total instances = 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 = Expected instances/total instances = (5C2)/(7C2) = 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 number of candies: Ʃ(P\*X) = Ʃ(1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120) = 3.090

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**

**Ans:**

|  | Mean | Median | Mode | Var | SD | Range | Comment |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Points | 3.596563 | 3.695 | 3.07, 3.92 | 0.2858814 | 0.534678736 | 2.17 | Not a normal distribution, data has skewness closer to 0 as data is equally distributed around mean and kurtosis is negative. Does not ave outliers. |
| Score | 3.21725 | 3.325 | 3.44 | 0.957379 | 0.978457443 | 3.911 | Somewhat normal distribution, will have positive skewness and kurtosis. Has outliers on upper extreme. |
| Weigh | 17.84875 | 17.71 | 17.02, 18.90 | 3.193166 | 1.786943236 | 8.4 | Somewhat normal distribution, will have positive skewness and kurtosis. Has one outlier on upper extreme. |

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?

Ans: Expected value = Ʃ(PX) = (108+110+123+134+135+145+167+187+199)/9 = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  | **Speed** | **Dist** |
| --- | --- | --- |
| **Skewness** | -0.117509861 | 0.80689496 |
| **Kurtosis** | -0.50899442 | 0.405052582 |
| **Inferences** | Closer to normal distribution, more values on right of mean, data is not concentrated around mean I.e not a peaked cureve, no outliers. | Closer to normal distribution, more values on left of mean, data is concentrated around mean I.e a peaked bell curve. 1 outlier on upper extreme. |

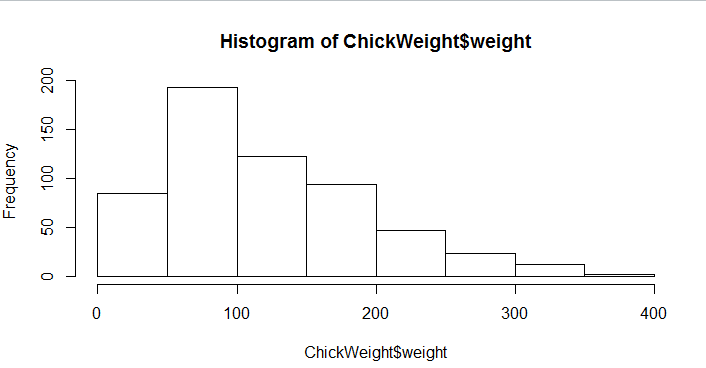
**SP and Weight(WT)**

**Use Q9\_b.csv**

Ans:

|  | SP | WT |
| --- | --- | --- |
| Skewness | 1.611450196 | -0.614753326 |
| Kurtosis | 2.977328944 | 0.950291491 |
| Inference | Closer to normal distribution, visual representation has longer tail to the right, it is highly peaked. Has outliers on upper extreme. | Not a normal distribution, Visual representation has longer tail to left. It not tat peaked. Has outliers on upper and lower extreme. |

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



Ans: Data is positively skewed as it has a longer tail on the right. It will have a negative kurtosis as data is distributed from 100-400, not very concentrated about mean. Mean of weight will range between 50-100.



Ans: Data has outliers on upper extreme. More values are concentrated near lower extreme which means data will have positive skewness and will have a longer tail on right also the kurtosis will be positive as data is concentrated near mean.

**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?

Mean of sample = 200 lbs, Std Dev of sample = 30 lbs

n = 2000

Av. Weight for 94% CI = stats.norm.interval(0.94, loc = 200, scale = 30/np.sqrt(2000)) = 198.7383, 201.2616

Av. Weight for 98% CI = stats.norm.interval(0.98, loc = 200, scale = 30/np.sqrt(2000)) = 198.4394,201.5606

Av. Weight for 96% CI = stats.norm.interval(0.96, loc = 200, scale = 30/np.sqrt(2000)) = 198.6223,201.3776

**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, Median = 40.5, variance = 25.529, std dev = 5.052

1. What can we say about the student marks?

Ans: Data is normally distributed, it is not concentrated near mean. Data is positively skewed.

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

Ans: If the distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness.

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

Ans:  If the mean is greater than the median, the distribution is positively skewed.

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

Ans:  If the mean is less than the median, the distribution is negatively skewed.

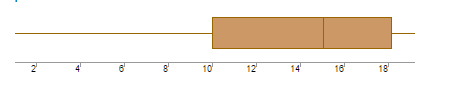
Q16) What does positive kurtosis value indicates for a data ?

Ans: Positive excess values of kurtosis (>3) indicate that a distribution is peaked and possess thick tails.

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

Negative excess values of kurtosis (<3) indicate that a distribution is flat and has thin tails.

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



What can we say about the distribution of the data?

Ans: Data is distributed asymmetrically.

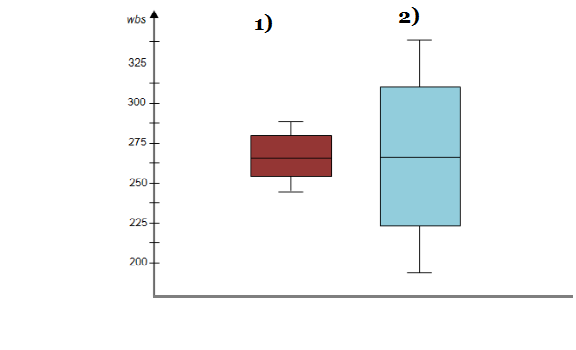
What is nature of skewness of the data?

Ans: Data is negatively skewed.

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

Ans: 8

Q19) Comment on the below Boxplot visualizations?



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

Ans: Median of both the data is 262.5. Boxplot 1 is more concentrated compared to boxplot 2. Both are normally distributed data. Both have zero outliers. Boxplot 1 will have positive kurtosis while boxplot 2 will have lower kurtosis. Skewness of both is 0.

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.

* 1. P(MPG>38) = 1-stats.norm.cdf(38, loc = 34.422, scale = 9.131) = 0.3475
  2. P(MPG<40) = stats.norm.cdf(40, loc = mean(cars.MPG), scale = stdev(cars.MPG)) = 0.7293

c. P (20<MPG<50) = stats.norm.cdf(50, loc = mean(cars.MPG), scale = stdev(cars.MPG)) - stats.norm.cdf(20, loc = mean(cars.MPG), scale = stdev(cars.MPG)) = 0.8988

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: As per the definition of normal distribution, the data should have a bell curve or median = mean = mode or shape is symmetric around mean. But mean(Cars.MPG) = 34.42, median(Cars.MPG)= 35.15, mode(Cars.MPG) = 29.62. Also data is not symmetric around mean. So this is somewhat similar to normal distribution.

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

Dataset: wc-at.csv

Ans: AT : Mean = 91.9, Median =90.8 ,Mode=94.5

Waist circumference : Mean = 101.89, Median = 96.54, Mode =121.0

As per the definition of normal distribution, the data should have a bell curve or median = mean = mode or shape is symmetric around mean. So this is not a normal distribution.

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

1. score(90%) = 1.2815

Z-score(94%) = 1.5547

Z-score(60%) = 0.2533

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

T(95%) = 1.7108

T(96%) = 1.8280

T(99%) = 2.4921

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: mean of sample = 260 days, stdev of sample = 90 days

Mu = 270 days, n = 18

H0:

T = (mean of sample - mu)/(stdev/sqrt(n)) = -0.4714

P = stats.t.cdf(-0.4714,17) = 0.3216 = 32.16%