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

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 | Interval |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Interval |

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

Ans. 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) 0

B) 1/6

C) 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?

Ans. 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 for a randomly selected child

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

= 3.09

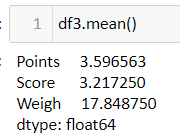
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset (Use Q7.csv)

* For Points, Score, Weight

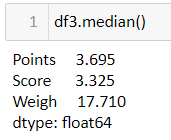
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Ans.**

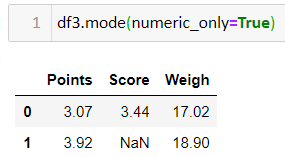
Mean:



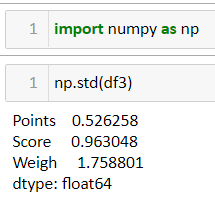
Median:



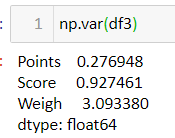
Mode:



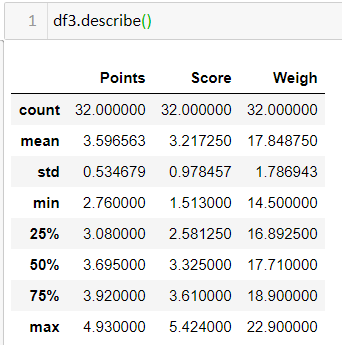
Standard Deviation:



Variance:



Some Inference based on the Q7.csv is



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. If we select randomly defining x = 108,110,123,134,135,145,167,187,199

Then using the python function random.choice(x) we can get any random value among these values defined in x.

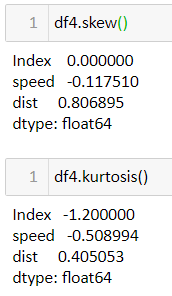
I tried it three times and the values are as follows:



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

**Car’s speed and distance (Use Q9\_a.csv)**

**Ans. Skewness and Kurtosis of the given Data set in Q9\_a.csv is as follows:**

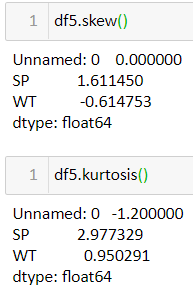
****

**On the basis of getting this outcome we can say that speed is negative skew data which also means a right-leaning curve whereas distance is positive skew which is a left-leaning curve.**

**Kurtosis is a measure of the thickness of the tail. In this we can see that speed is exceeding -1 which indicates that it is a very flat curve for speed. Overall if we say that speed graph is not normal, and distance data being less than +1 it is not so peaked data. So, distance values indicate a normal distribution data.**

**SP and Weight (WT)**

**Use Q9\_b.csv**

****

**On the basis of getting this outcome we can say that SP is positive skew data which also means a left-leaning curve whereas WT is negative skew which is a right-leaning curve.**

**Kurtosis is a measure of the thickness of the tail. In this we can see that SP is exceeding +1 which indicates that it is a very peaked curve. We say that SP graph is not normal, and WT data being less than +1 it is not so peaked data. So, WT values indicate a normal distribution data.**

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



Ans. Seeing the histogram, plotted above we can conclude that as the Chick Weight increases the Frequency decreases after reaching the highest point. The most common Chick Weight where Frequency is also high is at Chick Weight = 100 and Frequency = 200. The average of above data will be around 200 Chick Weight and 100 Frequency.

Seeing the box plot, we can say that the box plot is without the measurements so it’s a little difficult to say what point is what frequency. Seeing this I can get an idea that the box plot has some outliers to it on the top end of it. Seeing the outliers, we can draw the conclusion that the data has some uncertain increase or decrease to it which is causing this outcome to be outside the plot. Where if start seeing the plot, we see the minimum value at the lowest of the box plot and the maximum at the top end of the plot above that we have outliers. The difference between the 1st quartile and the mean value is not so much that is the reason this 1st quartile and mean are close to each other when compared with 3rd quartile and mean.

This is what we can conclude with the above data.

**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?

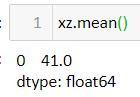
**Q12)** Below are the scores obtained by a student in tests

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

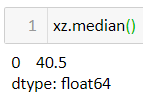
1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?

Ans. 1) Below are the Mean, Median, Variance, Standard Deviation of the above list:

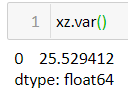
Mean:



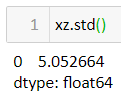
Median:



Variance:



Standard Deviation:



2) Seeing the mean and median which are 41 and 40.5 we can say that the students got average marks in the particular subject. Here we can also see that the mean > median which makes the distribution of the data as positively skewed.

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

Ans. If the mean median of a data is equal the skewness of the data is 0.

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

Ans. If the mean > median then the distribution is positively skewed.

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

Ans. If the median > mean then the distribution is negatively skewed.

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

Ans. A positive kurtosis value indicates the distribution is peaked and possess thick tail for a data. It has a higher peak and taller tail than a normal distribution.

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

Ans. A negative kurtosis value indicates the distribution is flat and has thin tail for a data. It has flatter peak with fewer values in its shorter tail as compared to a normal distribution.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

Ans. We can tell seeing this box plot that the median of this data is more toward to 3rd quartile which is around 15-16 range and the 3rd quartile is at around 18 and the 1st quartile is around 10. The maximum and the minimum of the data is not visible as we are not able to see the end of the plot. The interquartile range of this data is 18-10=8 approx.

For knowing the skewness of this data, we need to look at two factors first is where the median is placed at which in case of this distribution is at more towards the upper part of the box which means close to the 3rd quartile this shows the negatively skewed distribution of the data. Second, we need to see is whisker placement on the plot. As in this data we are not able to judge the whisker placement. Thus, we conclude that this distribution is negatively skewed.

Q19) Comment on the below Boxplot visualizations?



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

Ans The 1st Box plot is ranging from around 237.5 as the minimum point to around 287.5 as the maximum point. The 1st quartile of this box plot is starting from around 250, the median is at 262.5 (approx.) and 3rd quartile is at 280 (approx.). Seeing this plot, we can say the median is almost equal to 1st and 3rd quartile thus making its skewness equal to 0. The data ranging in the 1st box plot is very less thus making it a box plot with a short whisker or say short tail box plot.

The 2nd box plot is ranging from less than 200 (approx.) as the minimum point and above 325 (approx.) as the maximum point and 1st quartile is at 225 (approx.), median is at 262.5 (approx.) and third quartile is at 312.5 (approx.). This plot seems to have the same median as 1st box plot but the difference is at the 1st quartile and 3rd quartile. But the median is equal to 1st and 3rd so it also has the same skewness as 1st box plot which is equal to 0. The data ranging in this 2nd box plot is high making it a box plot with normal size whisker.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

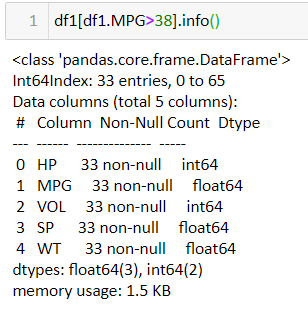
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

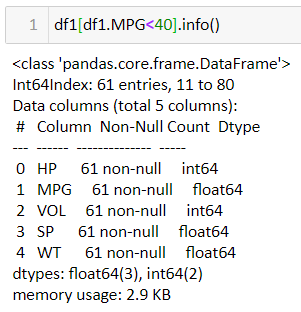
* 1. P(MPG>38)
  2. P(MPG<40)

c. P (20<MPG<50)

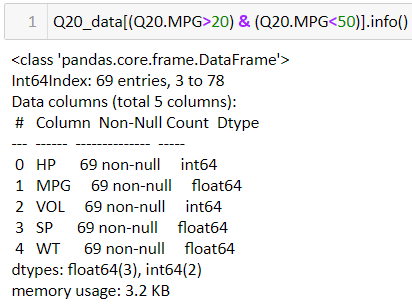
Ans. A) P(MPG>38) = 33/81



B) P(MPG<40) = 61/81



C) P(20<MPG<50) = 69/81



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

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 2.

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