|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| Weight of a person | Continue |
| 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 | Ordinal |
| 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 | Ratio |
| Years of Education | Ordinal |

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

S={HHH, HHT, HTH, HTT, THT, THH, TTH, TTT}

P(Two heads and one tail)=3/8

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

S={(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)}

1. Equal to 1 = 0
2. Less than or equal to 4 = 6/36
3. Sum is divisible by 2 and 3 = 25/36

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?

= (2C2+2C1\*3C1+3C2)/(7C2) = 6/7

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 |

=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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | points | score | Weight |
| Mean | 3.596563 | 3.21725 | 17.84875 |
| Median | 3.695 | 3.325 | 17.71 |
| variance | 0.285881 | 0.957379 | 3.193166 |
| standard deviation | 0.534679 | 0.978457 | 1.786943 |
| Range | 2.17 | 3.911 | 8.4 |
| Mode | 3.92 | 3.44 | 17.02 |

1. For all 3 variables, there is no much difference in Mean, median, and mode. Form this we can say that there are no outliers in the data.
2. Score and weight data is highly variated as compare to points data.

**Use Q7.csv file**

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 – (108+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**

**Use Q9\_a.csv**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | skew | Kurt | | speed | -0.11751 | -0.50899 | | Dist | 0.806895 | 0.405053 |   Speed –Skewness is very less, data is symmetric even the no kurtosis. So data distribution is similar to normal distribution.  Dist – Skewness vale is near to 1 indicate that data is slightly negatively skewed with small peak. |  |
|  |  |
|  |  |

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | skew | kurt |
| SP | 1.61145 | 2.977329 |
| WT | 0.61475 | 0.950291 |

SP – Skewness value is more than 1 indicate that data is highly negatively skewed. Positive kurtosis value more than 2 indicate that data is more peaked and with long thick tail.

WT –Positive skewness value more than 0.5 indicate that data is slightly negatively skewed and with positive kurtosis value saying data have small peak and long tail.

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

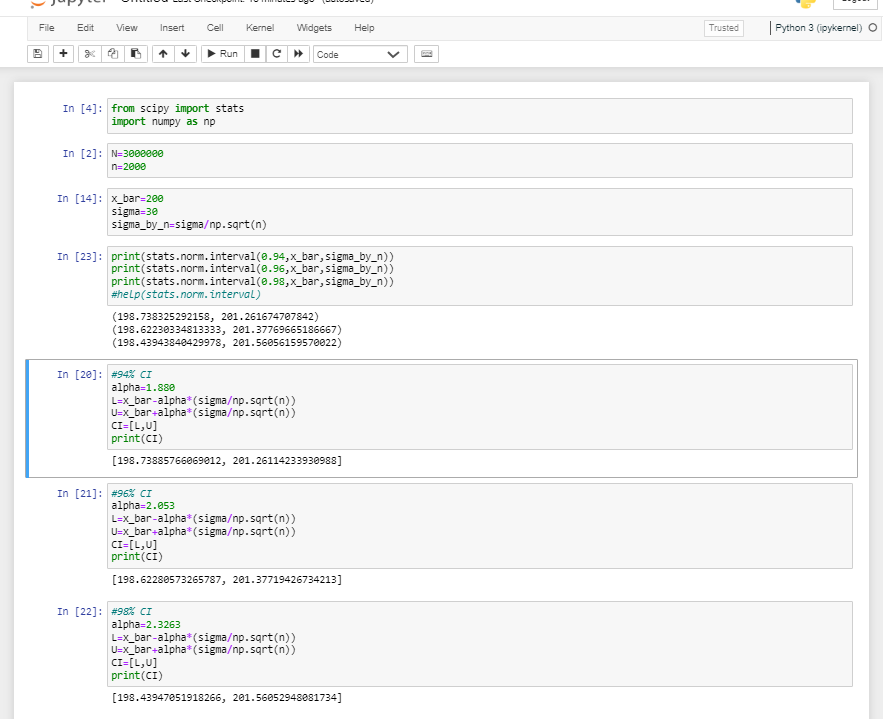


Data is positively skewed. High frequently chick weight is between 50 to 100 range.



Data have some upper outeliers. Data is positively skewed. Mean > median.

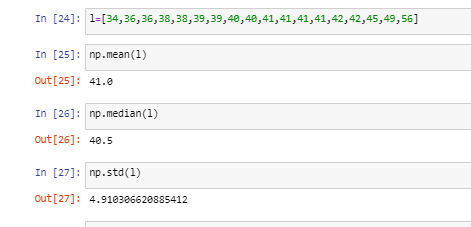
**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate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



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

Skewness is zero, data is normally distributed about the mean and median.

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

Data is negatively skewed.

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

Data is positively skewed.

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

Data is more peaked and have long thick tail.

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

No peak in data, almost flat distribution, thin and small tail.

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



What can we say about the distribution of the data?

50 % data lies between 10 to 18 value, median is approximately 15.

What is nature of skewness of the data?

* Data is negatively skewed.

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

Middle half of data -IQR  
10-18

Q19) Comment on the below Boxplot visualizations?



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

Boxplot 1 – Data lies between approximately 240 to 280 range (small range)

Boxplot 2 – Data lies between approximately 190 to 330 range (wide range)

Median value for both the data set is almost same value. Both the data sets 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

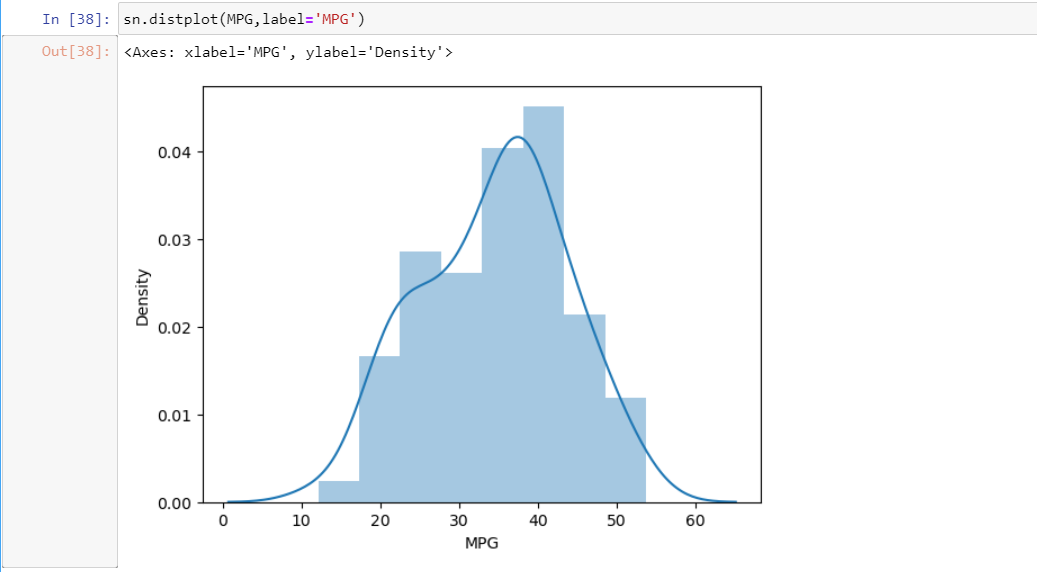
* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)



Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

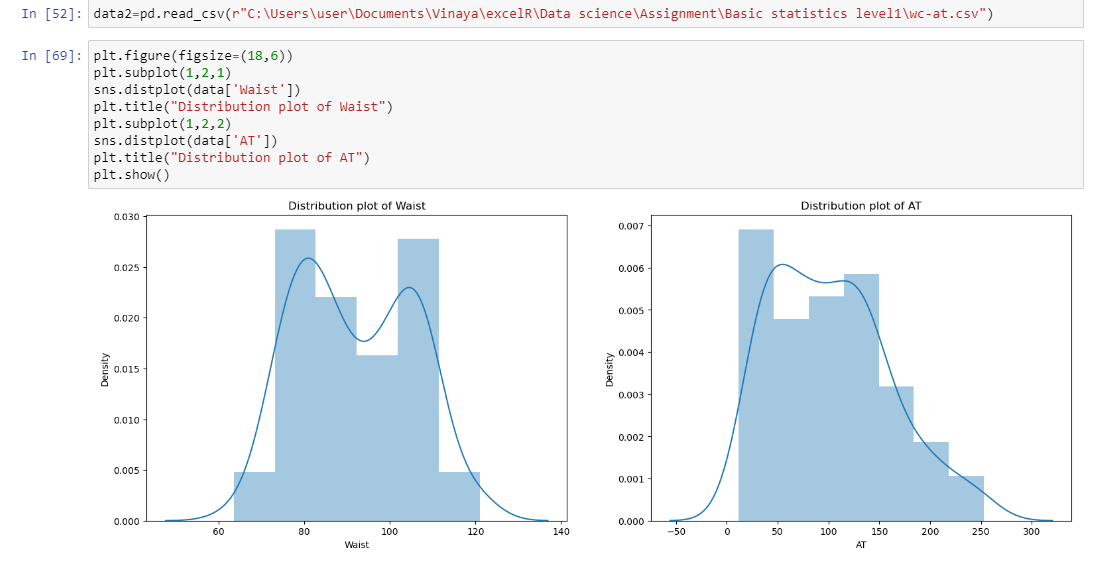
Dataset: Cars.csv



Slightly negatively skewed.

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

Dataset: wc-at.csv



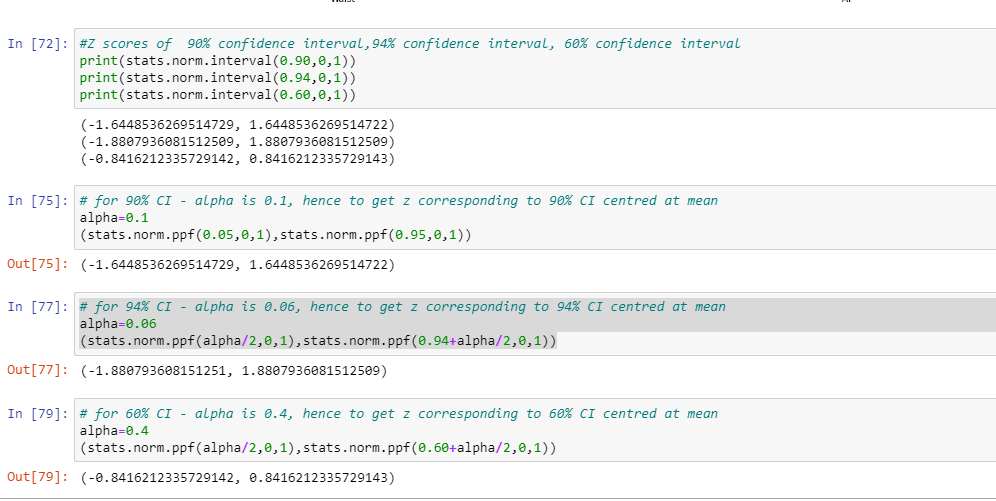
1. Waist circumference

* Waist data is slightly normally distributed having 2 peaks.

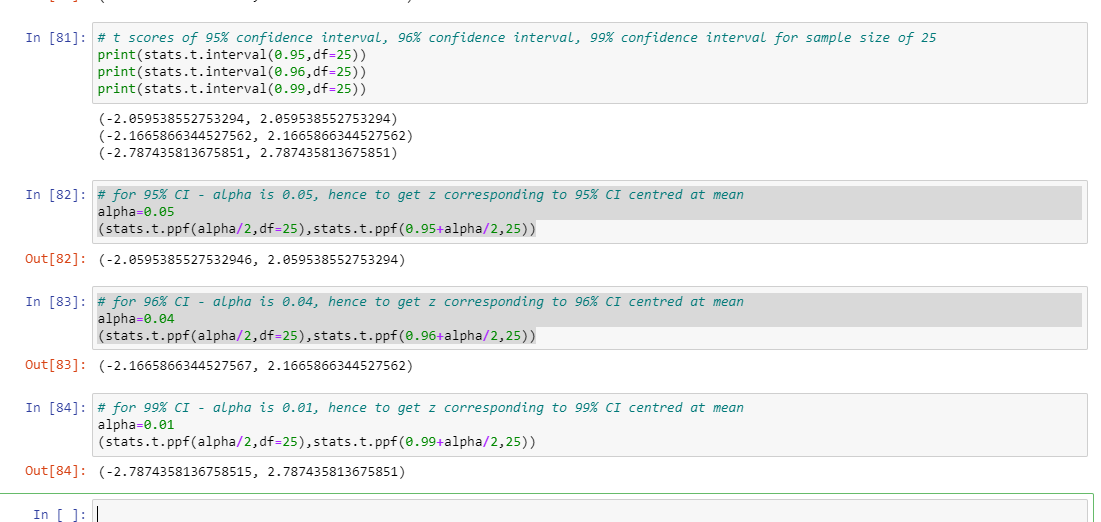
1. Adipose Tissue

* Right tailed, slightly positively distributed.

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 25



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

Population mean = mu = 270

Sample size = n = 18

Sample mean = x\_bar = 260

Sample standard deviation = s= 90

P(x\_bar < 260)

P((x\_bar-mu)/s/sqrt(n) < 260 – 270/90/sqrt(18))

P(t<-0.471)

0.3218

stats.t.cdf(-0.471,17)

0.3218140331685075