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

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

**Sol**- P (Head Head Tail) + P (Head Tail Head) + P (Tail Head Head)

Probability= number of favorable outcomes/Total number of outcomes

=1/8 + 1/8 + 1/8

= 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**-

1. If two dices were rolled, then total possible cases =36Total Favorable cases (Having sum =1) = 0As minimum sum is 2 for outcome (1,1).Hence, probability is = 0/36
2. (1,3), (2,2), (3,1), (1,1), (2,1), (1,2) = 3 outcomes 6/36 = 1/6
3. (1,5), (2,4), (3,3), (4,2), (5,1), (6,6) = 6 outcomes 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?

**Sol:** P (2Red, 3Green, 2Blue)

P (5/7)/P (5/3) =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:**

1×0.015+4×0.20+3×0.65+5×0.005+6×0.01+2×0.120=0.015+0.80+1.95+0.025+0.06+0.24=3.090So, the coming result is = 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**

**Sol:**

By using ‘Q7.csv’ file below I got the results for above question.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mazda RX4 | 3.9 | 2.62 | 16.46 |
| Mazda RX4 Wag | 3.9 | 2.875 | 17.02 |
| Datsun 710 | 3.85 | 2.32 | 18.61 |
| Hornet 4 Drive | 3.08 | 3.215 | 19.44 |
| Hornet Sportabout | 3.15 | 3.44 | 17.02 |
| Valiant | 2.76 | 3.46 | 20.22 |
| Duster 360 | 3.21 | 3.57 | 15.84 |
| Merc 240D | 3.69 | 3.19 | 20 |
| Merc 230 | 3.92 | 3.15 | 22.9 |
| Merc 280 | 3.92 | 3.44 | 18.3 |
| Merc 280C | 3.92 | 3.44 | 18.9 |
| Merc 450SE | 3.07 | 4.07 | 17.4 |
| Merc 450SL | 3.07 | 3.73 | 17.6 |
| Merc 450SLC | 3.07 | 3.78 | 18 |
| Cadillac Fleetwood | 2.93 | 5.25 | 17.98 |
| Lincoln Continental | 3 | 5.424 | 17.82 |
| Chrysler Imperial | 3.23 | 5.345 | 17.42 |
| Fiat 128 | 4.08 | 2.2 | 19.47 |
| Honda Civic | 4.93 | 1.615 | 18.52 |
| Toyota Corolla | 4.22 | 1.835 | 19.9 |
| Toyota Corona | 3.7 | 2.465 | 20.01 |
| Dodge Challenger | 2.76 | 3.52 | 16.87 |
| AMC Javelin | 3.15 | 3.435 | 17.3 |
| Camaro Z28 | 3.73 | 3.84 | 15.41 |
| Pontiac Firebird | 3.08 | 3.845 | 17.05 |
| Fiat X1-9 | 4.08 | 1.935 | 18.9 |
| Porsche 914-2 | 4.43 | 2.14 | 16.7 |
| Lotus Europa | 3.77 | 1.513 | 16.9 |
| Ford Pantera L | 4.22 | 3.17 | 14.5 |
| Ferrari Dino | 3.62 | 2.77 | 15.5 |
| Maserati Bora | 3.54 | 3.57 | 14.6 |
| Volvo 142E | 4.11 | 2.78 | 18.6 |

**For Points:**

* Mean = 3.596563
* Median = 3.695
* Mode = 3.07, 3.92
* Variance = 0.2858814
* Standard Deviation = 0.5346787
* Range = 2.76 4.93

**For Score:**

* Mean = 3.21725
* Median = 3.325
* Mode = 3.44
* Variance = 0.957379
* Standard Deviation = 0.9784574
* Range = 1.513 5.424

**For Weigh:**

* Mean = 17.84875
* Median = 17.71
* Mode =17.02, 18.90
* Variance = 3.193166
* Standard Deviation = 1.786943
* Range = 14.5 22.9

**Comments:**

1) **“Points”** and “**Score”** these two columns have mean and median close to each other but for “**Weigh”** it’s slightly different.

2) **“Points”** and “**Weigh”** are Bimodal.

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:**

Probability of selecting each patient = 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

**So, the expected Value of the Weight of that patient is = 145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Sol:**

By using the ‘**Q9\_a.csv’** data set I got the results below for above question

|  |  |  |
| --- | --- | --- |
| Index | speed | dist |
| 1 | 4 | 2 |
| 2 | 4 | 10 |
| 3 | 7 | 4 |
| 4 | 7 | 22 |
| 5 | 8 | 16 |
| 6 | 9 | 10 |
| 7 | 10 | 18 |
| 8 | 10 | 26 |
| 9 | 10 | 34 |
| 10 | 11 | 17 |
| 11 | 11 | 28 |
| 12 | 12 | 14 |
| 13 | 12 | 20 |
| 14 | 12 | 24 |
| 15 | 12 | 28 |
| 16 | 13 | 26 |
| 17 | 13 | 34 |
| 18 | 13 | 34 |
| 19 | 13 | 46 |
| 20 | 14 | 26 |
| 21 | 14 | 36 |
| 22 | 14 | 60 |
| 23 | 14 | 80 |
| 24 | 15 | 20 |
| 25 | 15 | 26 |
| 26 | 15 | 54 |
| 27 | 16 | 32 |
| 28 | 16 | 40 |
| 29 | 17 | 32 |
| 30 | 17 | 40 |
| 31 | 17 | 50 |
| 32 | 18 | 42 |
| 33 | 18 | 56 |
| 34 | 18 | 76 |
| 35 | 18 | 84 |
| 36 | 19 | 36 |
| 37 | 19 | 46 |
| 38 | 19 | 68 |
| 39 | 20 | 32 |
| 40 | 20 | 48 |
| 41 | 20 | 52 |
| 42 | 20 | 56 |
| 43 | 20 | 64 |
| 44 | 22 | 66 |
| 45 | 23 | 54 |
| 46 | 24 | 70 |
| 47 | 24 | 92 |
| 48 | 24 | 93 |
| 49 | 24 | 120 |
| 50 | 25 | 85 |

* Firstly, I exported this ‘csv data’ in ‘Jupyter Notebook’.
* Later I imported required libraries to read the dataset.
* To read this data I imported the ‘Pandas’ library as **‘import pandas as pd’**
* Before reading the ‘csv data’ successfully by using pandas library. First, I store the data in ‘df’ (you can give any name to store the data).
* df=pd.read\_csv("Q9\_a.csv”)
* Now we have to find the Skewness & Kurtosis values for speed and distance in cars dataset by using the ‘df.skew()’ & ‘df.kurt()’

Now the final values are:

* **Skewness** for **“speed”** = -0.11, skewness value is negative, so it is left skewed.
* **Kurtosis for “speed”** = -0.508, kurtosis value is less than Normal Kurtosis that is 3.
* **Skewness** for **“Distance”** = 0.80, skewness value is positive, so it is right skewed.
* **Kurtosis for “Distance”** = 0.405, kurtosis value is nearly equal to Normal Kurtosis that is 3.

**SP and Weight (WT)**

**Use Q9\_b.csv**

**Sol:**

By using the ‘**Q9\_b.csv’** data set I got the results below for above question.

|  |  |  |
| --- | --- | --- |
|  | SP | WT |
| 1 | 104.1854 | 28.76206 |
| 2 | 105.4613 | 30.46683 |
| 3 | 105.4613 | 30.1936 |
| 4 | 113.4613 | 30.63211 |
| 5 | 104.4613 | 29.88915 |
| 6 | 113.1854 | 29.59177 |
| 7 | 105.4613 | 30.30848 |
| 8 | 102.5985 | 15.84776 |
| 9 | 102.5985 | 16.35948 |
| 10 | 115.6452 | 30.92015 |
| 11 | 111.1854 | 29.36334 |
| 12 | 117.5985 | 15.75353 |
| 13 | 122.1051 | 32.81359 |
| 14 | 111.1854 | 29.37844 |
| 15 | 108.1854 | 29.34728 |
| 16 | 111.1854 | 29.60453 |
| 17 | 114.3693 | 29.53578 |
| 18 | 117.5985 | 16.19412 |
| 19 | 114.3693 | 29.92939 |
| 20 | 118.4729 | 33.51697 |
| 21 | 119.1051 | 32.32465 |
| 22 | 110.8408 | 34.90821 |
| 23 | 120.289 | 32.67583 |
| 24 | 113.8291 | 31.83712 |
| 25 | 119.1854 | 28.78173 |
| 26 | 114.5985 | 16.04317 |
| 27 | 120.7605 | 38.06282 |
| 28 | 119.1051 | 32.83507 |
| 29 | 99.56491 | 34.48321 |
| 30 | 121.8408 | 35.54936 |
| 31 | 113.4846 | 37.04235 |
| 32 | 112.289 | 33.23436 |
| 33 | 119.9211 | 31.38004 |
| 34 | 121.3926 | 37.57329 |
| 35 | 111.289 | 32.70164 |
| 36 | 115.0131 | 31.91122 |
| 37 | 114.0934 | 28.754 |
| 38 | 116.9094 | 27.87992 |
| 39 | 116.9094 | 28.6305 |
| 40 | 128.4613 | 30.11543 |
| 41 | 116.3926 | 37.39252 |
| 42 | 115.7488 | 35.02718 |
| 43 | 117.4613 | 30.52743 |
| 44 | 114.0934 | 28.34398 |
| 45 | 114.381 | 33.07863 |
| 46 | 117.1051 | 32.62192 |
| 47 | 118.2087 | 36.49862 |
| 48 | 116.4729 | 33.91006 |
| 49 | 127.9094 | 28.0706 |
| 50 | 118.289 | 33.45847 |
| 51 | 118.289 | 33.21395 |
| 52 | 118.289 | 33.43671 |
| 53 | 120.4043 | 40.39816 |
| 54 | 143.3926 | 37.62069 |
| 55 | 135.3926 | 37.25439 |
| 56 | 126.4043 | 40.58907 |
| 57 | 110.4613 | 30.14754 |
| 58 | 118.289 | 32.73452 |
| 59 | 112.6452 | 30.61528 |
| 60 | 115.5766 | 37.66287 |
| 61 | 130.2087 | 36.88815 |
| 62 | 117.6685 | 37.86041 |
| 63 | 126.0481 | 43.39099 |
| 64 | 125.3123 | 40.72283 |
| 65 | 128.1284 | 40.15948 |
| 66 | 126.5985 | 15.71286 |
| 67 | 132.4846 | 37.97996 |
| 68 | 133.6802 | 41.57397 |
| 69 | 133.3123 | 40.47204 |
| 70 | 158.3007 | 37.14173 |
| 71 | 164.5985 | 15.82306 |
| 72 | 133.416 | 44.01314 |
| 73 | 133.1401 | 43.35312 |
| 74 | 124.7152 | 52.99775 |
| 75 | 121.8642 | 42.6187 |
| 76 | 132.8642 | 42.77822 |
| 77 | 169.5985 | 16.13295 |
| 78 | 150.5766 | 37.92311 |
| 79 | 151.5985 | 15.76963 |
| 80 | 167.9445 | 39.4231 |
| 81 | 139.8408 | 34.94861 |

* Firstly, I exported this ‘csv data’ in ‘Jupyter Notebook’.
* Later I imported required libraries to read the dataset.
* To read this data I imported the ‘Pandas’ library as **‘import pandas as pd’**
* Before reading the ‘csv data’ successfully by using pandas library. First, I store the data in ‘dx’ (you can give any name to store the data).
* dx=pd.read\_csv("Q9\_b.csv”)
* Now we have to find the Skewness & Kurtosis values for speed and weight by using the ‘dx.skew()’ & ‘dx.kurt()’

Now the final values are:

* **Skewness** for **“SP”** = 1.611, skewness value is positive, so it is right skewed.
* **Kurtosis for “SP”** = 2.97, kurtosis value is close to Normal Kurtosis that is 3, so SP is normal.
* **Skewness** for **“Weight”** = -0.61, skewness value is negative, so it is left skewed.
* **Kurtosis for “Weight”** = 0.95, kurtosis value is less than Normal Kurtosis that is 3.

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



**Sol:**

**For Histogram:**

* The most of datapoints are in the range 50-100 with frequency 200.
* And least range of weight is 400 some were around 0-10.
* Skewness- we can notice a long tail towards right so it is heavily right skewed.



**Sol:**

* Median is less than mean right skewed and we have outlier on the upper side of box plot and there is less data points between Q1 and bottom point.

**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?

Sol:

* The **information given**is: Sample meanof X 200
* Sample standard deviation of S =30
* Sample size n = 2000
* X+/-(t1\*S/sqrt(n))
* Degree of freedom = 2000-1 = 1999
* Considering a **94%** confidence level

**T1 = 1.56**

**= 200+/-(1.56\*30/ sqrt of 2000)**

**= (197.06, 202.93)**

* Considering a **96%** confidence level

T1 = 1.75

= **200+/-(1.75\*30/ sqrt of 2000)**

**= (196.40, 203.60)**

* Considering a **98%** confidence level

T1 = 2.05

**= 200+/-(2.05\*30/ sqrt of 2000)**

**= (195.22, 205.77)**

**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?

**Sol:**

1. **Find mean, median, variance, standard deviation.**

* Mean = 41,
* Median = 40.5,
* Variance = 25.52941
* Standard Deviation = 5.052664

1. **What can we say about the student marks?**

* we can say that mean and median are closer to each other, seems no outlier present.

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

Sol:

* symmetrical, equally oriented

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

Sol:

* Mean > Median: Positively skewed (right-skewed) distribution.

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

Sol:

* Median > Mean: Negatively skewed (left-skewed) distribution.

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

Sol:

* The data is sharper and has long tails

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

Sol:

* The data is not Sharper and under the normal distribution with short tail.

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)?   
  
**Sol:**

**What can we say about the distribution of the data?**

* Most of observations are having value above 10 and observations whose value is above 15 are 40%.

**What is nature of skewness of the data?**

* Left skewed, median is greater than mean.

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

* Approximately= -8

Q19) Comment on the below Boxplot visualizations?



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

Sol:

* By observing both the plots the whiskers level is same from the Q1 and Q3 to their respective upper and lower limit and again mean and median are also equal so we can say distribution is symmetrical.

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)

**Sol:**

From ‘Cars.csv’ data set we have



MPG <- Cars$MPG

1. P(MPG>38)

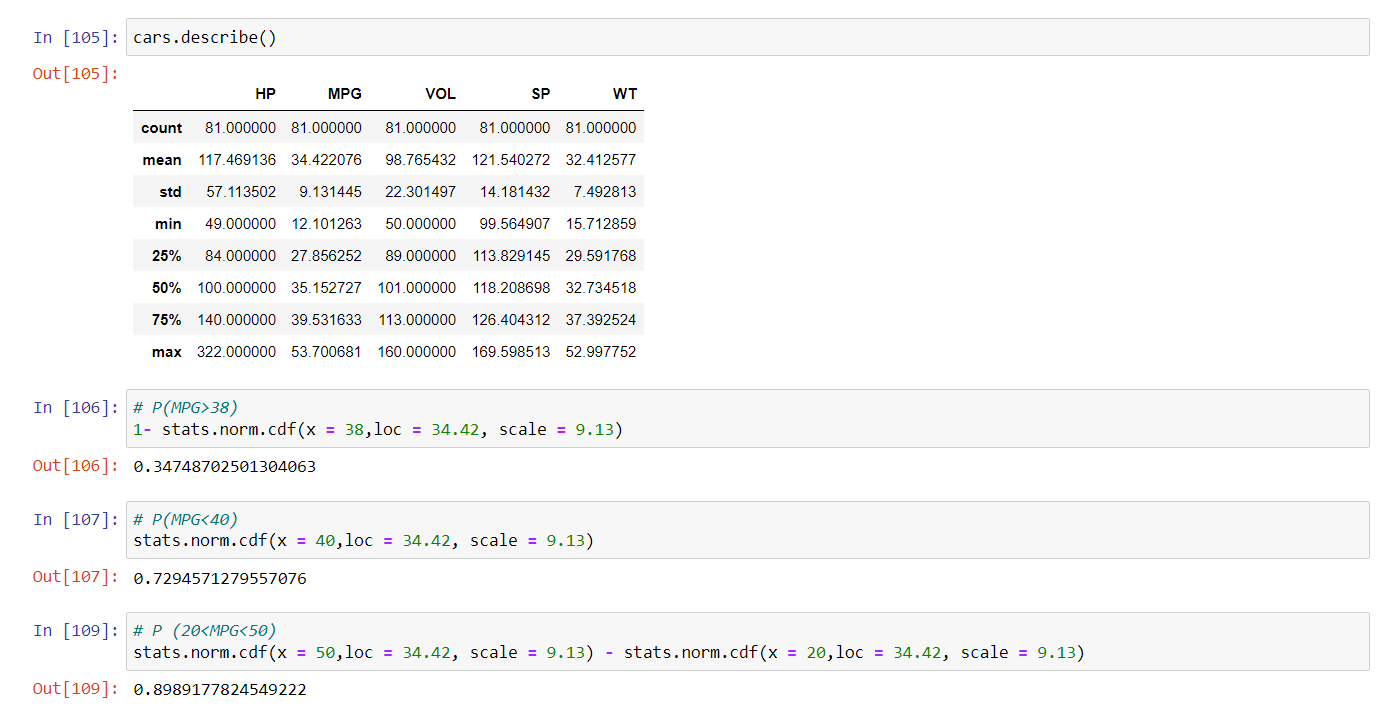
0.34

1. P(MPG<40)

0.72

1. P (20<MPG<50)

0.89



Q 21) Check whether the data follows normal distribution

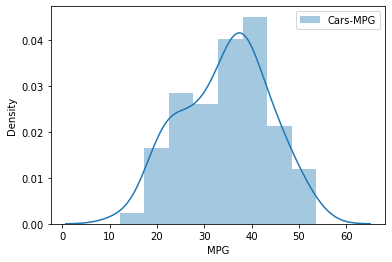
1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Sol:

**The MPG of cars not following Normal Distribution.**

**Skewness = -0.177**

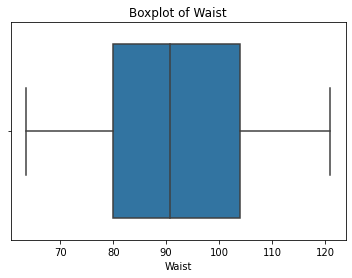
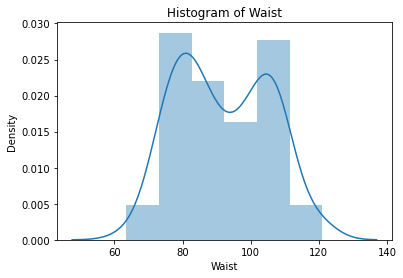


1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

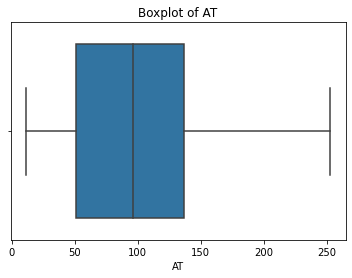
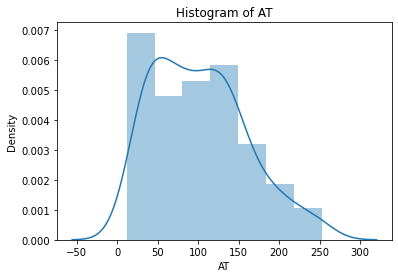
Dataset: wc-at.csv

Sol:

**Mean greater than median, both the whisker is of same length, median is slightly shifted towards left. Data is fairly symmetric.**



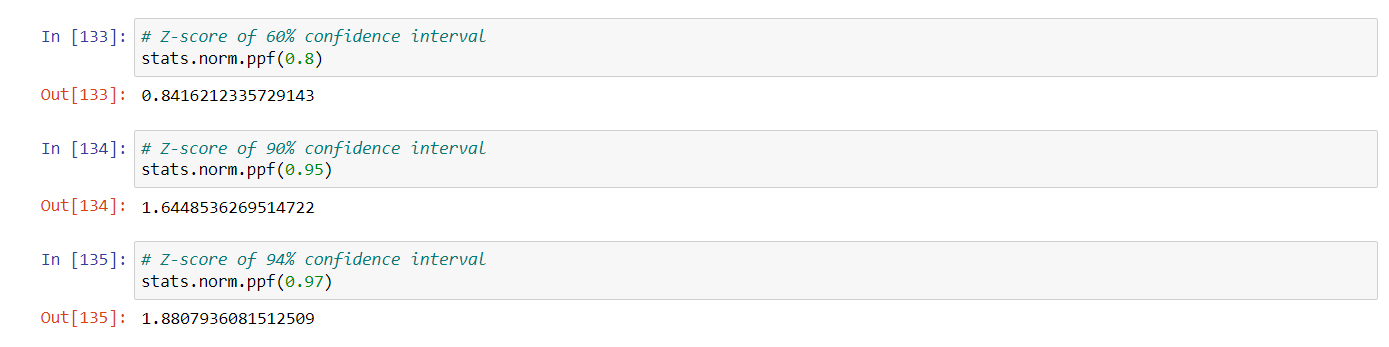
**Mean greater than median, right whisker is larger than left whisker, data is positively skewed.**



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

Sol:

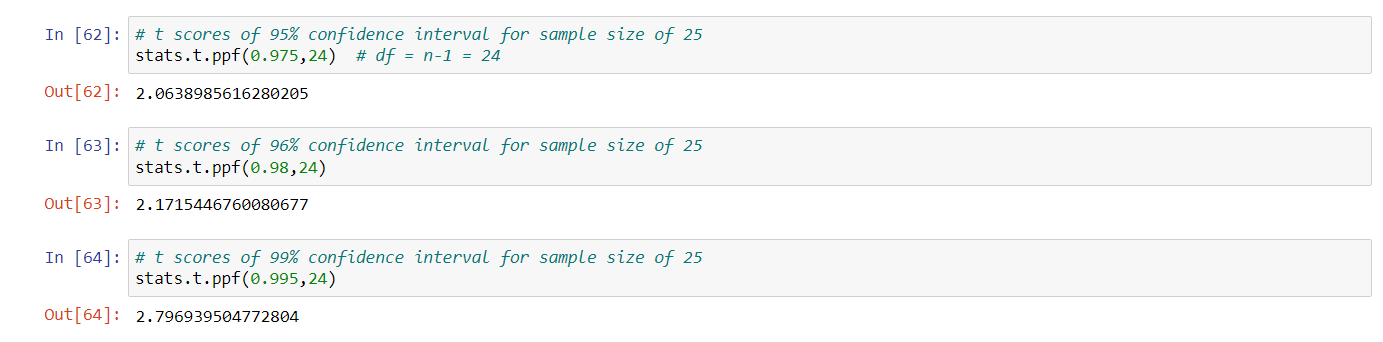
* **Z score of 60% Confidence Interval = 0.841**
* **Z score of 90% Confidence Interval = 1.644**
* **Z score of 94% Confidence Interval = 1.880**



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

**Sol:**

|  |  |
| --- | --- |
| **Confidence Interval** | **T Score** |
| **95%** | **2.06** |
| **96%** | **2.17** |
| **99%** | **2.79** |



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

**Sol:**

**Population mean = 270 days**

**Sample mean = 260 days**

**Sample SD = 90 days**

**Sample n = 18 bulbs**

**df = n-1 = 17**

**t = {(260-270) / (90/√18)}**

**t = (-1 \* √2) / 3**

**t = - 0.471**

**For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.**

