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

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

**Answer: Solution**

Let S be the Sample space then

S = {H,T}, {H,T}, {H,T} then the , event E is {HHT, HTH, THH}

Let X be a random variable denoting the two heads and one tail.

P(X = 2) = Probability of occurrence of 2 heads and 1 tail.

= P(HHT) + P (HTH) + P (THH)

= 1/2.1/2.1/2 + 1/2.1/2.1/2 + 1/2.1/2.1/2 = 3/8.

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

1. Equal to 1

**Answer:** If two dice were rolled, then total possible cases = 36

Total Favorable cases (Having sum = 1) = 0

As minimum sum is 2 for outcome (1, 1).

**Hence, probability is 0**

1. Less than or equal to 4
2. Sum is divisible by 2 and 3

**Solution:**

You are using two ‘fair’ six-sided dice (faces numbered 1–6).  
Throwing one die gives 6 possible outcomes Throwing two dice gives 6\*6 = 36 possible outcomes, with sums between 2 (double 1) and 12 (double 6). The only numbers between 2–12 which are divisible by both 2 and 3 are 6 and 12.  
While 12 can only be made 1 way (double 6) 6 can be made 5 ways (1, 5 2, 4 3, 3 4, 2 5, 1) so the probability of throwing a sum of either 12 or 6 is (number of successful outcomes)/ (number of all possible outcomes) = (1+5)/36 = 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?

**Answer: Solution**

Total number of balls  
= (2 + 3 + 2)   
= 7  
Let S be the sample space  
Then, n(S) = Number of ways of drawing 2 balls out of 7

n(S)=7C2

n(S) =

n(S) =21  
Let E = Event of 2 balls, none of which is blue  
∴ n(E) = Number of ways of drawing 2 balls out of (2 + 3) balls

n(E) =

n(E) =

n(E) =10

∴P(E) =

**=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

**Answer: Solution**

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

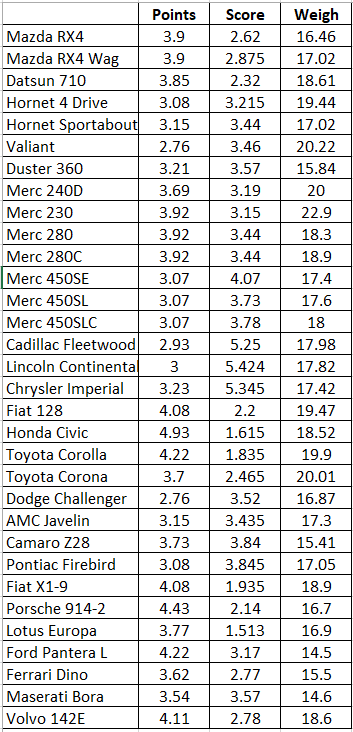
**Expected number of candies for a randomly selected child = 3.09**

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**



**Answer: Solution (Points)**

1. **Mean**

= (3.9+3.9+3.85+3.08+3.15+2.76+3.21+3.69+3.92+3.92+3.92+3.07+3.07+3.07+2.93+3+3.23+4.08+4.93+4.22+3.7+2.76+3.15+3.73+3.08+4.08+4.43+3.77+4.22+3.62+3.54+4.11) = 115.09

**= 115.09/32 = 3.59**

1. **Median**

**= 3.695**

1. **Mode**

= 3.92, 3.07

1. **Variance**

= (3.9+3.9+3.85+3.08+3.15+2.76+3.21+3.69+3.92+3.92+3.92+3.07+3.07+3.07+2.93+3+3.23+4.08+4.93+4.22+3.7+2.76+3.15+3.73+3.08+4.08+4.43+3.77+4.22+3.62+3.54+4.11) = 115.09

Total Square is115.09 x 115.09 = 13245.708100000002

and then divided the number of items in the data set 13245.708100000002 / 32 =413.92837812500005

Squaring up all the numbers from the data set individually and adding them:

(2.76\*2.76) + (2.76\*2.76) + (2.93\*2.93) + (3\*3) + (3.07\*3.07) + (3.07\*3.07) +(3.07\*3.07) + (3.08\*3.08) + (3.08\*3.08) + (3.15\*3.15) + (3.15\*3.15) + (3.21\*3.21)+ (3.23\*3.23) + (3.54\*3.54) + (3.62\*3.62) + (3.69\*3.69) + (3.7\*3.7) + (3.73\*3.73)+ (3.77\*3.77) + (3.85\*3.85) + (3.9\*3.9) + (3.9\*3.9) + (3.92\*3.92) + (3.92\*3.92) +(3.92\*3.92) + (4.08\*4.08) + (4.08\*4.08) + (4.11\*4.11) + (4.22\*4.22) + (4.22\*4.22)+ (4.43\*4.43) + (4.93\*4.93) = 422.7907

Performing subtraction of the individual squared and the total squared 422.7907 - 413.92837812500005 = 8.862321874999964Number of items in data set 32Dividing the subtraction value within i.e., 328.862321874999964 / 32 = 0.27694755859374887

**Variance = 0.27694755859374887**

1. **Standard Deviation**

=

=

=

= 0.27694755859375

= √0.27694755859375

**= 0.52625807223619**

1. **Range**
2. = Largest value – Smallest Value
3. = 4.93- 2.76
4. **= 2.17**

**Note: Calculation Procedure same for Point**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weight |
| Mean | 3.59 | 3.217 | 17.849 |
| Median | 3.695 | 3.325 | 17.710 |
| Mode | 3.92, 3.07 | 3.440 | 17.020 |
| Variance | 0.276 | 0.957 | 3.139 |
| Standard Deviation | 0.526 | 0.978 | 1.787 |
| Range | 2.147 | 3.911 | 8.4 |

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?

**Answer: Solution**

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)

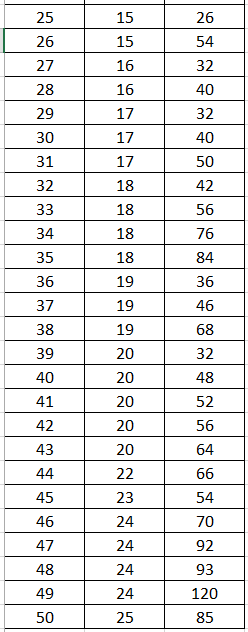
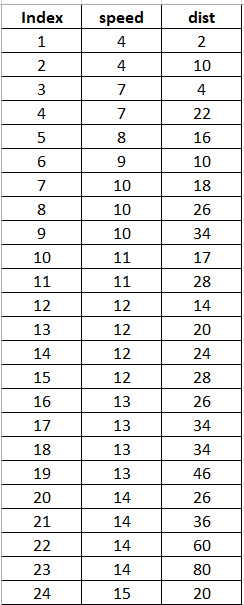
**= 145.33**

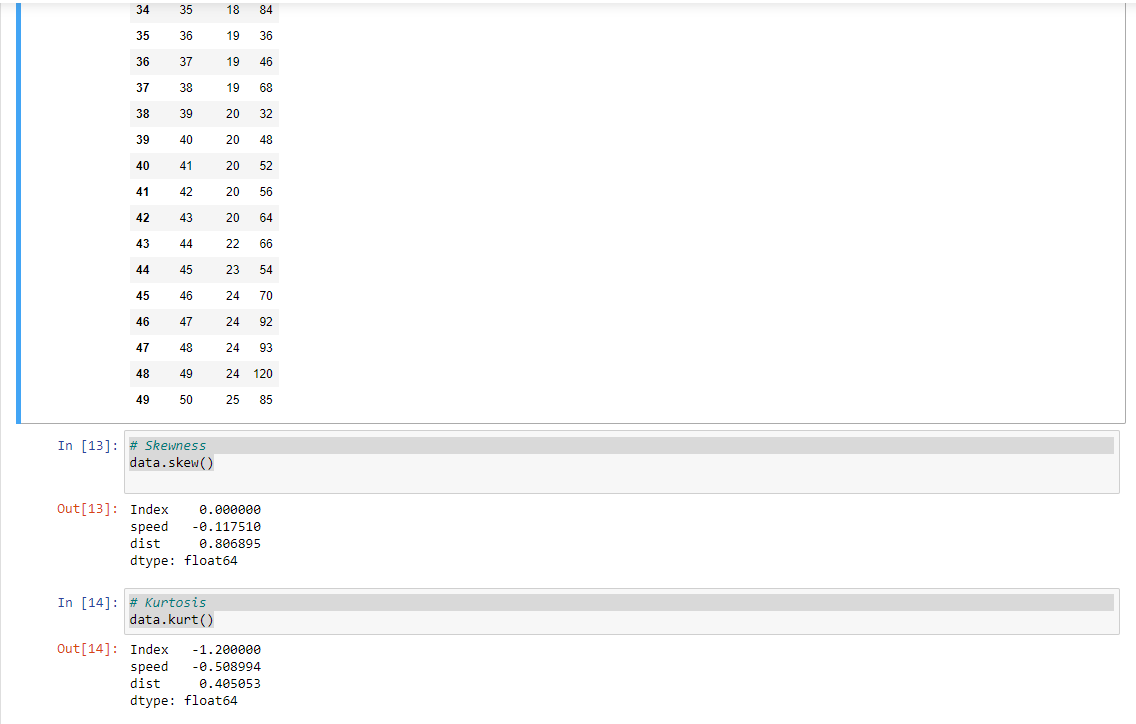
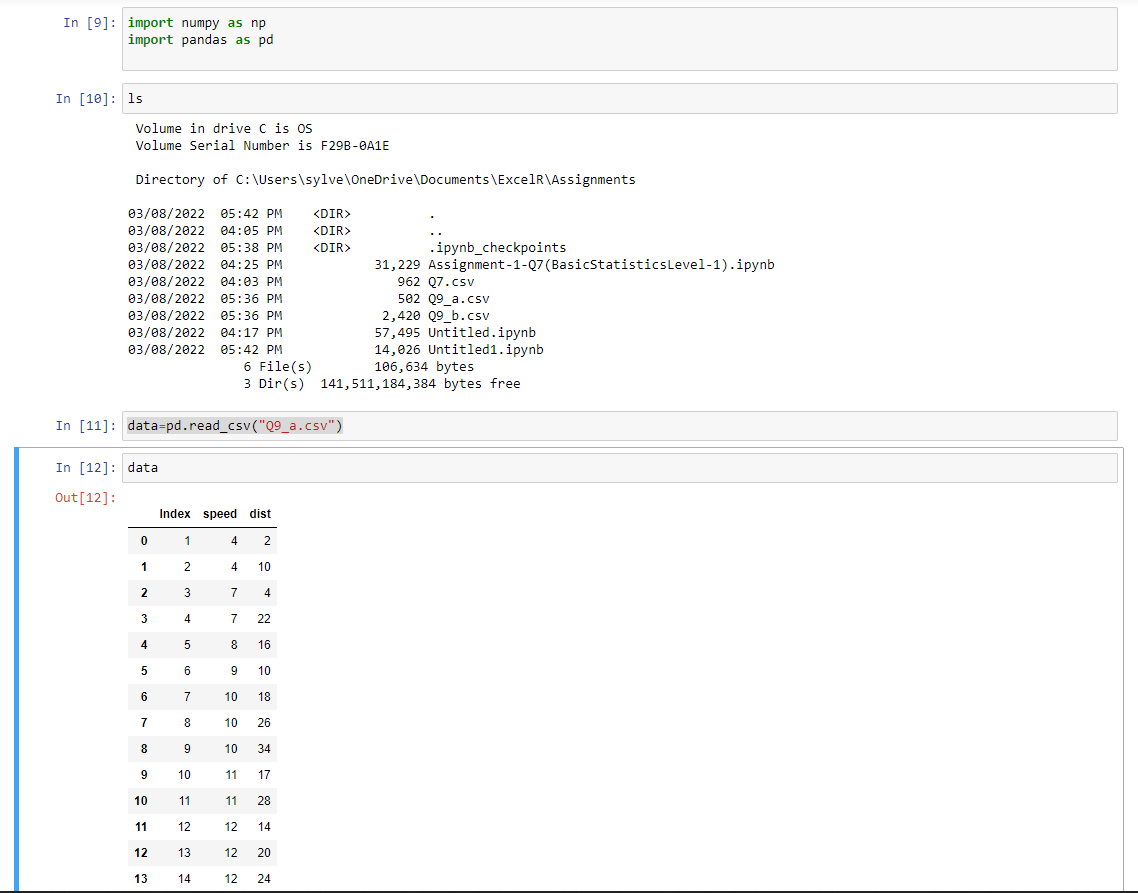
**Expected Value of the Weight of that patient = 145.33**

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

**Cars speed and distance**

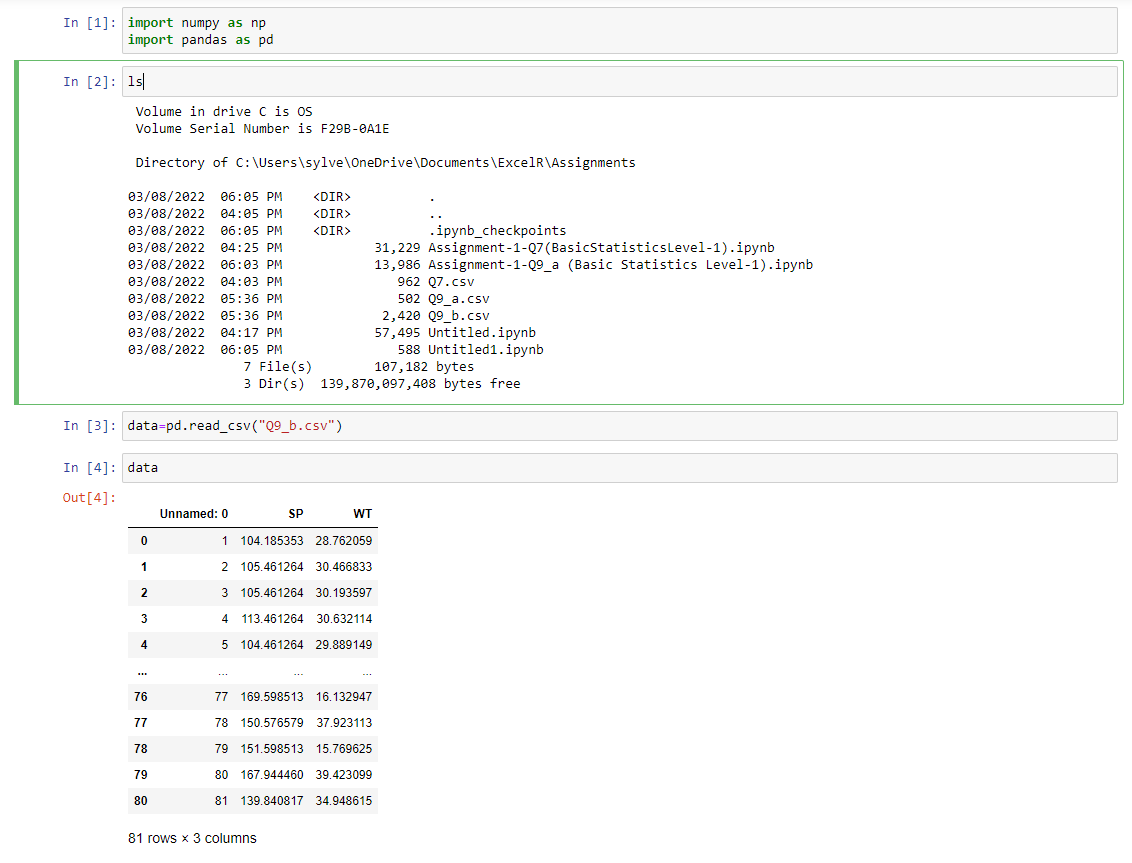
**Use Q9\_a.csv**



**SP and Weight (WT)**

**Use Q9\_b.csv**





Skewness for speed= -0.1139548, skewness value is negative so it is left skewed. Since magnitude is slightly greater than 0 it is slightly left skewed

And for distance= 0.7824835, right skewed (Positive) slight magnitude to right

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



**Solution:** The most of the data points are concerted in the range 50-100 with frequency 200.

And least range of weight is 400 somewhere around 0-10.

So, the expected value the above distribution is 75.

Skewness- we can notice a long tail towards right so it is heavily right skewed.



**Solution**:

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?

**Answer: Solution**

Using the t-distribution, it is found that:

* The 94% confidence interval is (198.73, 201.27).
* The 96% confidence interval is (198.61, 201.39).
* The 98% confidence interval is (198.43, 201.57).

We are given the standard deviation for the sample, which is why the t-distribution is used to solve this question.

The information given is:

* Sample mean of https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%3D%20200.
* Sample standard deviation of https://tex.z-dn.net/?f=s%20%3D%2030.
* Sample size of https://tex.z-dn.net/?f=n%20%3D%202000.

The interval is:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%5Cpm%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D

* In which t is the critical value for the two-tailed confidence interval.

Considering a 94% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 1.8916, hence:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20-%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20-%201.8916%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20198.73

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%2B%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20%2B%201.8916%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20201.27

The 94% confidence interval is (198.73, 201.27).

Considering a 96% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 2.0673, hence:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20-%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20-%202.0673%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20198.61

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%2B%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20%2B%202.0673%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20201.39

The 96% confidence interval is (198.61, 201.39).

Considering a 98% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 2.3452, hence:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20-%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20-%202.3452%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20198.43

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%2B%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20%2B%202.3452%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20201.57

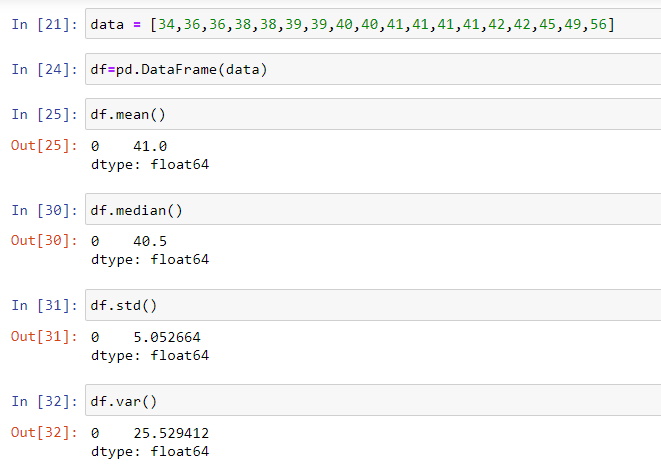
The 98% confidence interval is (198.43, 201.57).

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, variance= 24.111, Standard deviation= 4.910.



1. What can we say about the student marks?

**Solution:**

There is no outlier because Mean > median. It is slightly skewed for Right side. So, there is no outlier

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

**Solution:** There is no outlier. Symmetrically distributed

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

**Solution:** Right Skewed. Positive direction.

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

**Solution:** left skewed. Negative direction.

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

**Solution:** The data is normally distributed, kurtosis value is 0.

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

**Solution:** The distribution of the data has lighter tails and a flatter peak than the normal distribution.

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



What can we say about the distribution of the data?

Let’s assume above box plot is about ages of the students in a school.

50% of the people are above 10 yrs. old and remaining are less.

And students whose age is above 15 are approx. 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)?

**Solution:** (Q3-Q1) =18-10=8

Q19) Comment on the below Boxplot visualizations?



**Draw an Inference from the distribution of data for Boxplot 1 w.r.t Boxplot 2.**

Boxplot 1 and Boxplot 2 “Median” is equal.

Boxplot 1 and Boxplot 2 are symmetric.

There is don’t have any outlier. By observing both the plots whisker’s level is high in boxplot 2, mean and median are equal hence distribution is symmetrical

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

Data \_set: Cars.csv

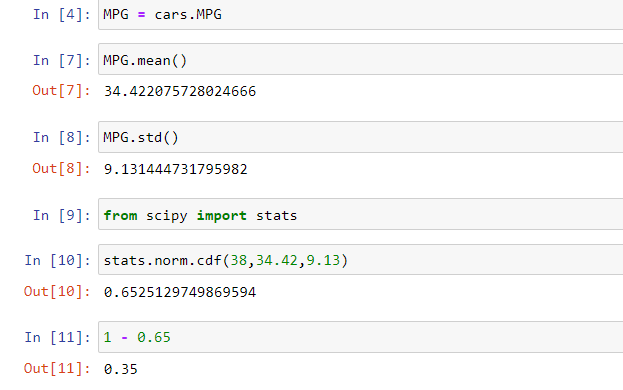
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

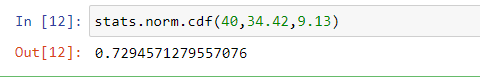
* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

**Solution:** z =

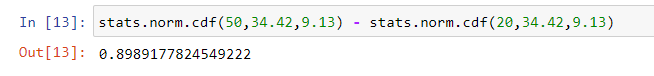
1. **P(MPG>38)**



1. **P(MPG<40)**



1. P (20<MPG<50)

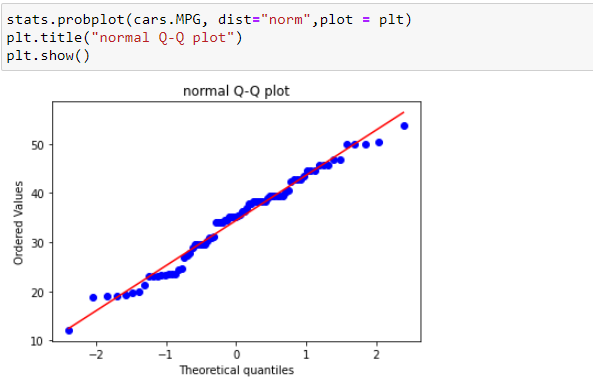


Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

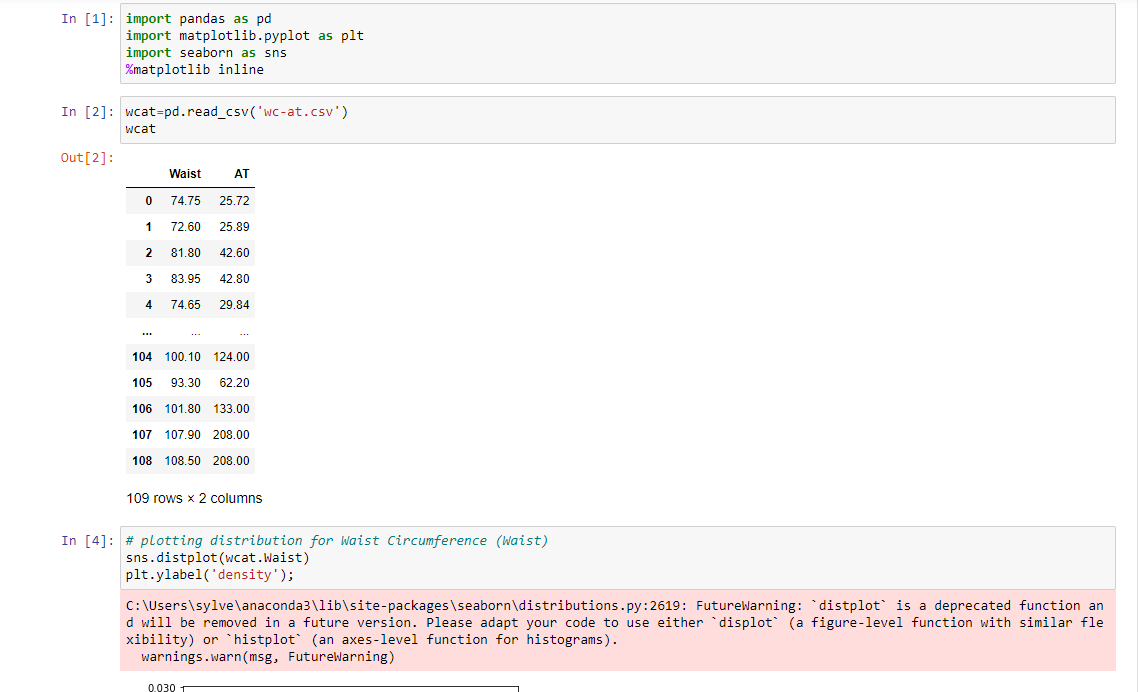
Dataset: Cars.csv

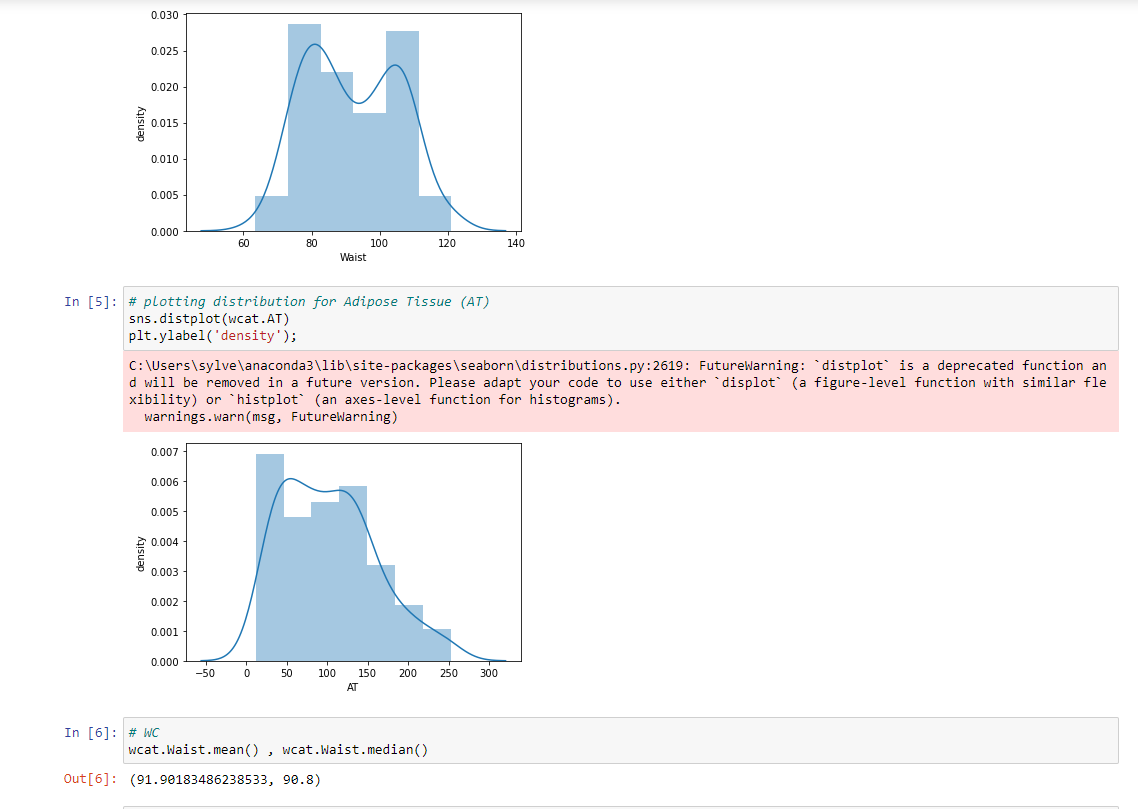
Solution: Normal distribution indicates in Q-Q plot

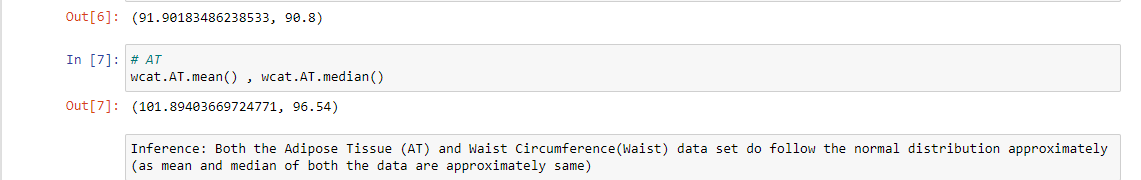


1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

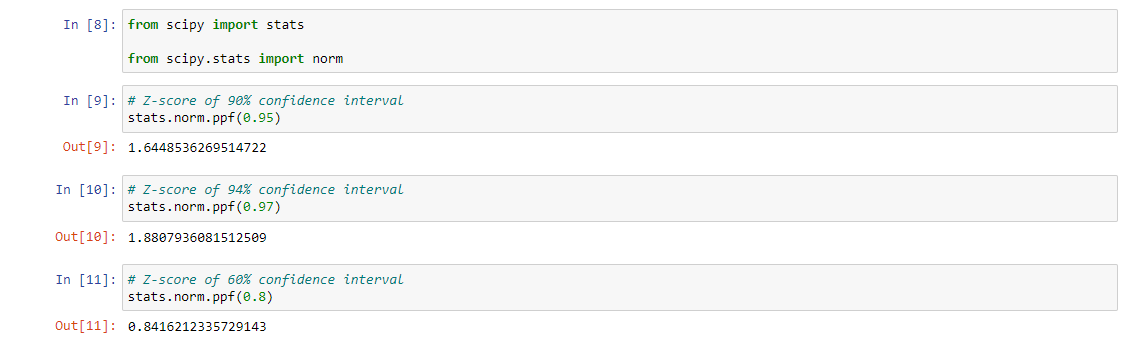
Dataset: wc-at.csv





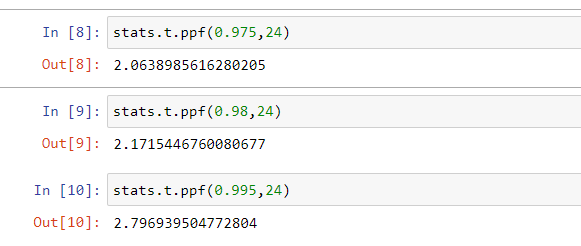


Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval



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

**Solution:** Confidence interval in 95%, 96%, 99%.



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

**To find:**

**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**

**Solution:**

t - Statistics for the data is given as follows: t=x-µ÷s/√n

x = mean of the sample of bulbs = 260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

t=230-270÷90/√18

t=-10÷90/3√2

t= -10÷30/√2

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.

The probability that t < - 0.471 with 17 degrees of freedom assuming the population mean is true, the t-value is less than the t-value obtained with 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of 0.3218 assuming the mean life of the bulbs is 300 days.

(or)

**Solution:**

µ=270,

=260,

SD=90,

n=18,

df=n-1=18-1= 17

T-score = = = -10/21.23

Graphical user interface, text

Description automatically generated = -0.47

**Probability = 32%**