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

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

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

**Answer :** 3 coins are tossed, so 2\*2\*2=8 possible outcomes.

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

Let t2 be event of getting 3 heads and one tail.

Hence, t2 = {HHT, HTH, THH}= 3/8

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

As when 2 dice are rolled the sample space = 6\*6 = 36 possibilities.

1. Equal to 1 --> **Answer** : 0

The minimum possible sum is (1,1)=2

Hence, probability of sum = 1 is 0/36 = 0

1. Less than or equal to 4 --> **Answer** : 6/36 = 1/6

Possible outcomes are {(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)}

Hence, probability of sum <= 4 is 6/36 = 1/6

1. Sum is divisible by 2 and 3 --> **Answer** : 5/36

Possible outcomes are {(1,5), (2,4), (4,2), (3,3), (5,1)}

Hence, probability of sum is divisible by 2 and 3 is 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?

**Answer -->** Sample Space={2red, 3green, 2blue}=7

Number of ways of drawing 2 balls out of 7 = 21

Number of ways of drawing 2 balls, none of which is blue. = 10

Probability that none of the balls drawn is blue = 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 -->** 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

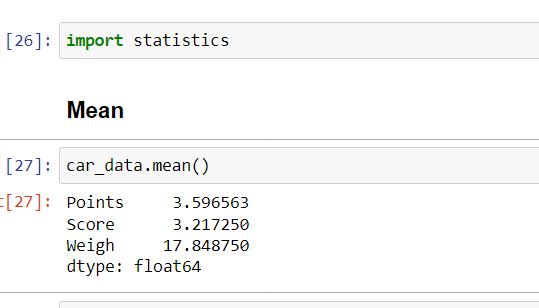
=3.09

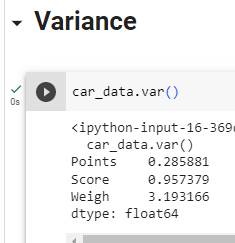
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given data set

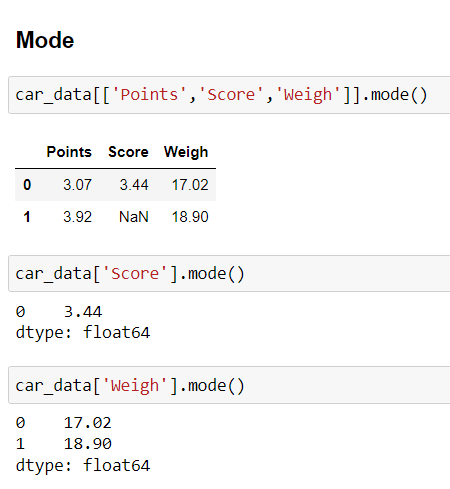
* For Points,Score,Weigh>

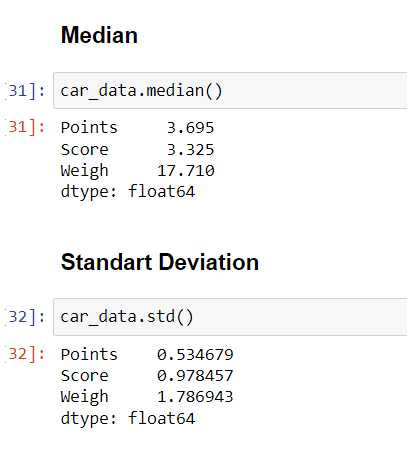
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

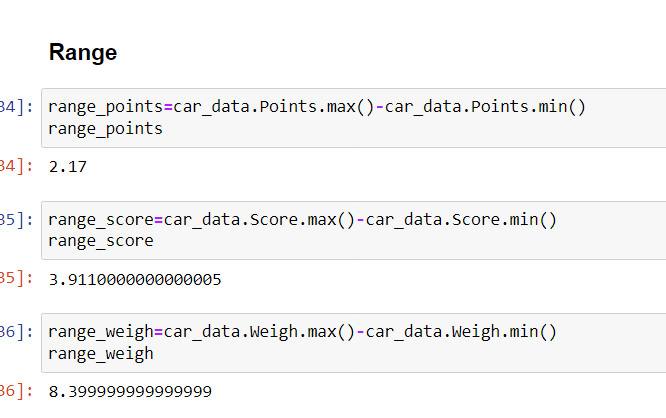
**Use Q7.csv file**

**Answer :** 

****







From above values we can conclude that the highest values among Points, Score and Weigh, Weigh carries higher values while among 3 columns Points carries lower values.

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 :** Expected value = Sum (X \* Probability of X)

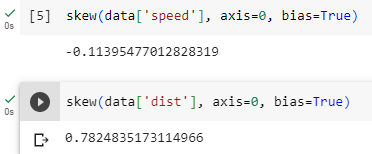
= (1/9)(108)+ (1/9)(110)+ (1/9)(123)+ (1/9)(134)+ (1/9)(145)+ (1/9)(167)+ (1/9)(187)+ (1/9)(199)

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

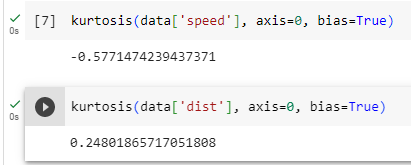
**Answer : **

The skewness of car speed is :- -0.11395477.

From this we can say that speed of the car data set is negatively skewed, that is tail is on the left side of the distribution. That is mean is lesser that median.

The skewness of car distance is :- 0.78248352.

From this we can say that distance of the car data set is positively skewed, that is tail is on the right side of the distribution. That is mean is greater that median.



The kurtosis of car speed is :- -0.57714742

From this we can say that speed of the car data set is having thinner tail than the normal distribution.

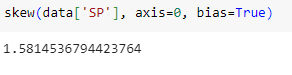
The kurtosis of car distance is :- 0.24801866

From this we can say that speed of the car data set is having thinner tail than the normal distribution.

**SP and Weight(WT)**

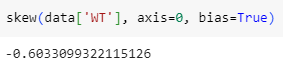
**Use Q9\_b.csv**

**Answer :** The skewness of SP is:- 1.5814536794423764



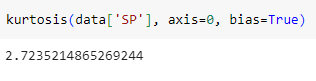
From this we can say that the data is positively skewed and the distribution has a long right tail. That is mean is greater that median.

The skewness of WT is:- -0.6033099322115126



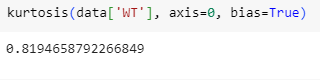
From this we can say that the data is negatively skewed and the distribution has a long left tail. That is mean is lesser that median.

The Kurtosis of SP is:- 2.7235214865269244



Kurtosis of SP has long tails, consequence of the outliers.

The Kurtosis of WT is:- 0.8194658792266849



Kurtosis of WT has long tails, consequence of the outliers. Also WT feature has bimodal distribution.

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



From the above histogram we can infer that, data is not normally distributed, instead distribution is positively skewed. Also we can infer that maximum weight is in between 50 to 100. Here, mean is greater that median.



From the above box plot we can infer that there are outliers present. From the whiskers we can infer that data is positive skewed.

**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 :** Sample – 2000, Population – 3000000, Sample Mean (x) – 200, Sample Std. Deviation (s) – 30

As we don’t have population std. deviation hence we will calculate CI through t-score:

n = 2000 🡪 n-1 = 1999

Code: stats.norm.interval(0.94, 200, 30/(2000\*\*0.5))

CI 94% 🡪 (198.738325292158, 201.261674707842)

Code: stats.norm.interval(0.98, 200, 30/(2000\*\*0.5))

CI 98% 🡪 (198.43943840429978, 201.56056159570022)

Code: stats.norm.interval(0.96, 200, 30/(2000\*\*0.5))

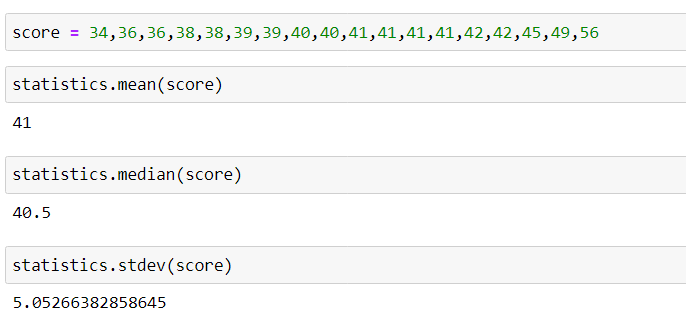
CI 96% 🡪 (198.62230334813333, 201.37769665186667)

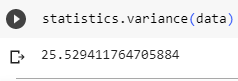
**Q12)** Below are the scores obtained by a student in tests

**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.

**Answer : Mean 🡪** 41 , **Median 🡪** 40.5 , **Variance 🡪** 25.529411764705884, **Standard Deviation 🡪** 5.05266382858645.





1. What can we say about the student marks?

**Answer :** From above data of students marks we can say that the maximum students got marks in between range of 35 and 45. Also there is only one student who scored 56 marks, that is the highest score. Also mean of marks of student is 41 which is slightly greater than median.

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

**Answer :** When mean and median of data are equal then density curve has no skew, that is zero skewness. The distribution would be the symmetrical distribution.

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

**Answer:** When mean > median, then distribution is positively skewed or right skewed distribution. That is distribution has long tail on its right side.

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

**Answer:** When mean < median, then distribution is negatively skewed or left skewed distribution. That is the distribution has long tail on its left side.

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

**Answer:** Positive kurtosis value indicates that the distribution has heavier tails than the normal distribution.

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

**Answer:** Negative kurtosis value indicates that the distribution has lighter tails than the normal distribution.

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



What can we say about the distribution of the data?

**Answer:** the data is distributed on the high end of the scale.

What is nature of skewness of the data?

**Answer:** data is left skewed data.

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

**Answer:** 18(Upper Quartile) – 10(Lower Quartile)

Q19) Comment on the below Boxplot visualizations?



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

**Answer:** There are not any outliers present in both the boxplots. Both the boxplots is showing symmetric distribution with median around 260.

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)

**Answer:** code 🡪 1 - stats.norm.cdf(38, 34.422076, 9.131445)

0.34759394041453007

* 1. P(MPG<40)

**Answer:** code 🡪 1 - stats.norm.cdf(40, 34.422076, 9.131445)

0.7293498604157946

* 1. P (20<MPG<50)

**Answer:** code 🡪 1 – [stats.norm.cdf(20, 34.422076, 9.131445) + [1 - stats.norm.cdf(50, 34.422076, 9.131445)]]

0.89886891

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

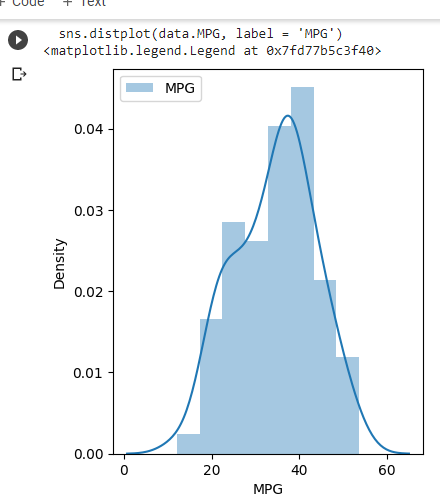
**Answer:** code🡪 plt.figure(figsize = (4,5))

sns.distplot(data.MPG, label = 'MPG')

plt.xlabel('MPG')

plt.ylabel('Density')

plt.legend(loc = 'upper left')



The distribution of MPG of cars is fairly normally distributed.

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

Dataset: wc-at.csv

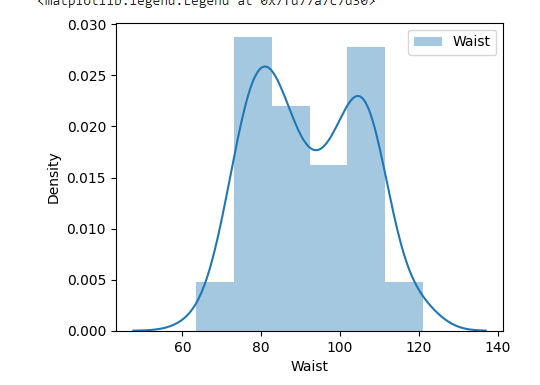
**Answer:** code🡪 plt.figure(figsize = (5,4))

sns.distplot(adipose.Waist, label = 'Waist')

plt.xlabel('Waist')

plt.ylabel('Density')

plt.legend(loc = 'upper right')



Waist Circumference has 2 peaks, data is fairly symmetrical.

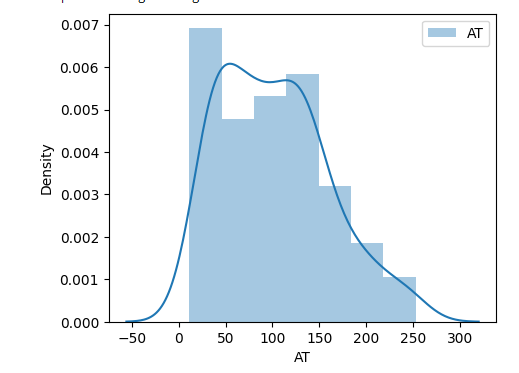
Code🡪 plt.figure(figsize = (5,4))

sns.distplot(adipose.AT, label = 'AT')

plt.xlabel('AT')

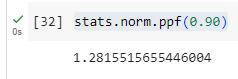
plt.ylabel('Density')

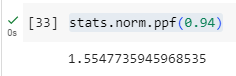
plt.legend(loc = 'upper right')

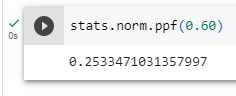


Data is slightly positively skewed.

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

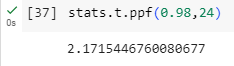
**Answer:** 90% confidence interval 🡪 

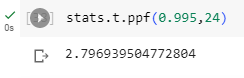
94% confidence interval 🡪 

60% confidence interval 🡪 

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

**Answer:** 95% confidence interval 🡪 

96% confidence interval 🡪 

99% confidence interval 🡪

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

Alternate hypothesis: average life of bulb < 260

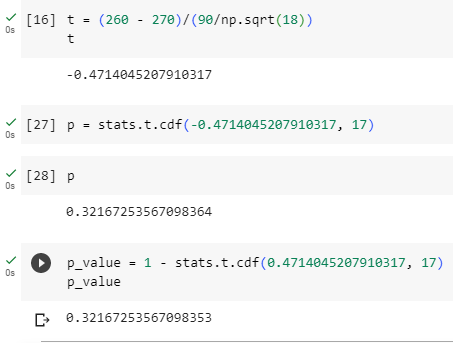
Null hypothesis: average life of bulb >= 260

Population mean 🡪 270;

Sample Mean 🡪 260;

Sample Std 🡪 90;

Sample size (n) 🡪 18



The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.32 or 32%.