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

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 | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

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

Solu:

Number of outcomes is 2^3 = 8.

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

event= two heads and one tail

P(A)={ THH, HTH, HHT}

Probability of two heads and one tail P(A)=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

Solu:

When 2 dice are rolled total possible outcomes are N(S)= 6^2=36

1. Equal to one:

Not possible sum is always exceeds 1: (1+1)

P(A)=0/36=0.

1. Less than or equal to 4

favorable outcome (sum less than or equal to 4) = { (1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}

P(B)=6/36=1/6

P(B)=1/6

C) Sum is divisible by 2 and 3 ={ (1 , 5) , (3 , 3) , (4 , 2) , (5 , 1) , (6 , 6) }

P(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?

Solu:

Total number of balls = (2 + 3 + 2)= 7

When first ball is drawn (except blue) = 5/7

When second ball is drawn (remaining balls and except blue)= 4/6=2/3

P(A)= (5/7)\*(2/3) = 10/21

P(A)=10/21

(OR)

Then, n(S) = Number of ways of drawing 2 balls out of 7

n(S)=7^C2

n(S)=(7×6)/(2×1)

n(S)=21

Let A = Event of 2 balls, none of which is blue

n(A) = Number of ways of drawing 2 balls out of (2 + 3) balls

n(A)=5C2n(E)=(5×4)/(2×1)

n(A)=10

P(A)=n(A)/n(S)=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

Solu:

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

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, Weight>

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

**Use Q7.csv file**

**By using python:**

**Mean:**

data.mean()

Points 3.61

Score 3.22

Weigh 17.85

**Median:**

**data.median()**

Points 3.695

Score 3.325

Weigh 17.710

**Mode:**

print('Mode of Points:', sts.mode(data['Points']))

print('Mode of Score:', sts.mode(data['Score']))

print('Mode of Weigh:', sts.mode(data['Weigh']))

Mode of Points: 3.92

Mode of Score: 3.44

Mode of Weigh: 17.02

**Variance:**

np.var(data)

Points 0.276

Score 0.927

Weigh 3.093

**Standard deviation:**

np.std(data)

Points 0.526

Score 0.963

Weigh 1.759

**Range:**

print('points',data['Points'].max()-data['Points'].min())

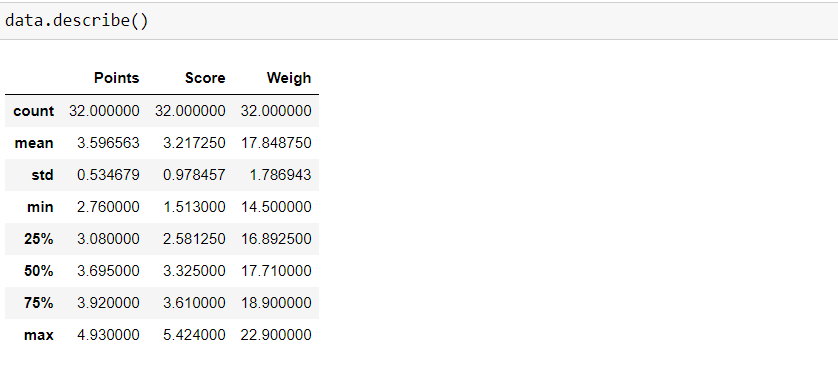
print('score',data['Score'].max()-data['Score'].min())

print('weigh',data['Weigh'].max()-data['Weigh'].min())

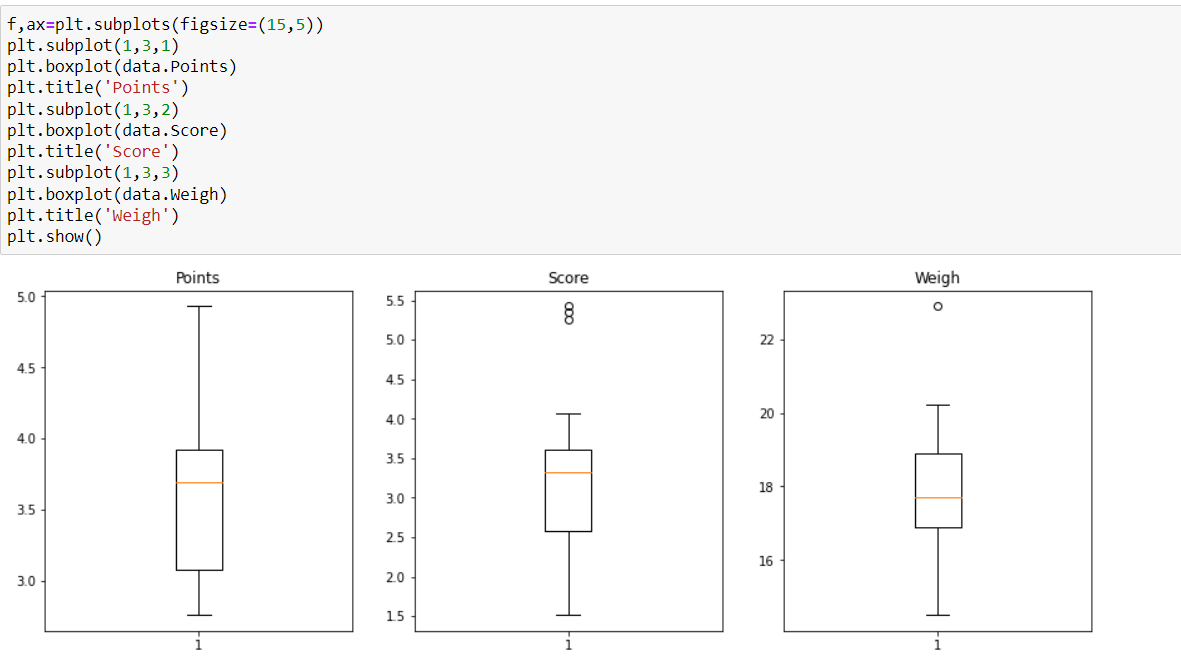
points 2.17

score 3.911

weigh 8.411



**Draw some inferences**



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?

Solu:

By using python

a= [108, 110, 123, 134, 135, 145, 167, 187, 199]

sum(a)/len(a)

=145.33

Expected value= (1/9) \* (1308)= 145.33

Expected Value of the Weight of random 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**

Solu:

If the skewness is between -0.5 & 0.5, the data are nearly symmetrical.

If the skewness is between -1 & -0.5 (negative skewed) or between 0.5 & 1(positive skewed), the data are slightly skewed.

If the skewness is lower than -1 (negative skewed) or greater than 1 (positive skewed), the data are extremely skewed.

**Skewness** for **Q9\_a.csv**

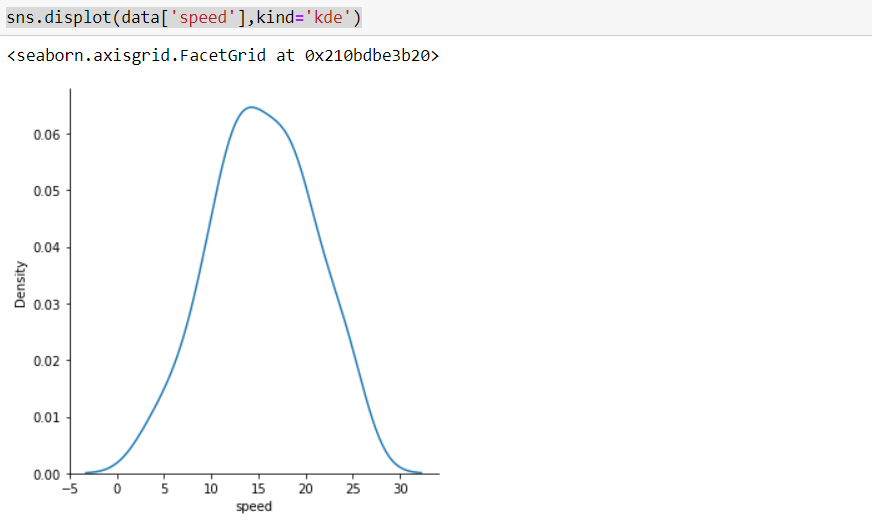
round(data['speed'].skew(),4) = -0.1175 (the data are nearly symmetrical.)

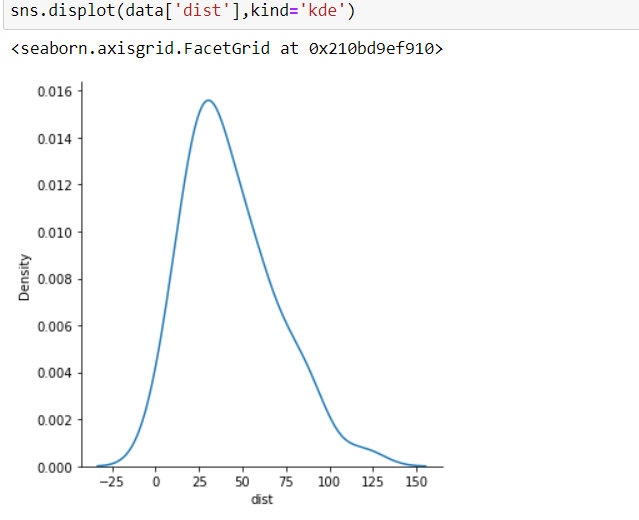
round(data['dist'].skew(),4)=0.8069 (positive skewed), the data are slightly skewed.

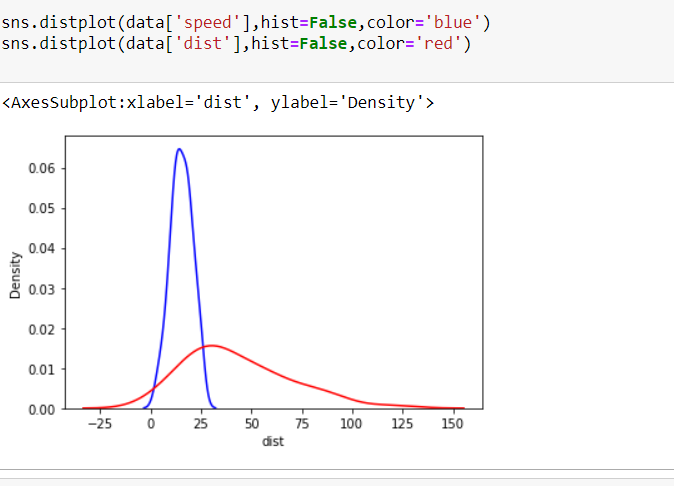
**Kurtosis**

round(data['speed'].kurtosis(),4)= -0.509

round(data['dist'].kurtosis(),4)= 0.4051







**Skewness** for **Q9\_b.csv**

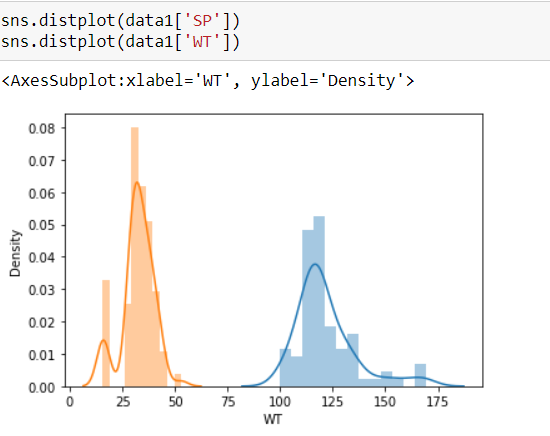
round(data1['SP'].skew(),4)=1.6115 (positive skewed), the data are extremely skewed.

round(data1['WT'].skew(),4)= -0.6148 (negative skewed)

**Kurtosis**

round(data1['SP'].kurtosis(),4)=2.9776

round(data1['WT'].kurtosis(),4)=0.9503



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



Ans: The histograms peak has right skew and tail is on right. Mean > Median>Mode (Positive skewness)

We have outliers on the higher side.

Ans: The boxplot has outliers on the maximum side.



**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?

Ans:

conf\_94 =stats.t.interval(alpha = 0.94, df=1999, loc=200, scale=30)

print(np.round(conf\_94,0))

conf\_98 =stats.t.interval(alpha = 0.98, df=1999, loc=200, scale=30)

print(np.round(conf\_98,0))

conf\_96 =stats.t.interval(alpha = 0.96, df=1999, loc=200, scale=30)

print(np.round(conf\_96,0))

For 94% confidence interval Range is= [143.54-256.46]

For 98% confidence interval range is [130.15-269.85]

For 96% confidence interval range is [138.35-261.65]

**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?

Solu:

Mean:41

Median:40.5

Variance:25.52

Standard deviation: 5.05

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

Ans: No skewness is present we have a perfect symmetrical distribution

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

Ans: Skewness and tail is towards Right , it is positive skewed

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

Ans: Skewness and tail is towards left, it is negative skewed

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

Ans: Positive kurtosis means the curve is more peaked and it is Leptokurtic

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

Ans:Platykurtic, Negative Kurtosis means the curve will be flatter and broader

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



What can we say about the distribution of the data?

Ans:In this boxplot , data are not normaly distributed, the median is in higher range

What is nature of skewness of the data?

Ans: The data is a skewed towards left. The whisker range of minimum value is greater than maximum.

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

Ans: Inter Quartile Range = upper quartile(Q3) – lower quartile (Q1)=18-10=8

Q19) Comment on the below Boxplot visualizations?



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

Ans: First there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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

a.P(MPG>38)

1-round(stats.norm.cdf(38,loc=data.MPG.mean(),scale=data.MPG.std()),3)

= 0.348

b.P(MPG<40)

round(stats.norm.cdf(40,loc=data.MPG.mean(),scale=data.MPG.std()),3) = 0.729

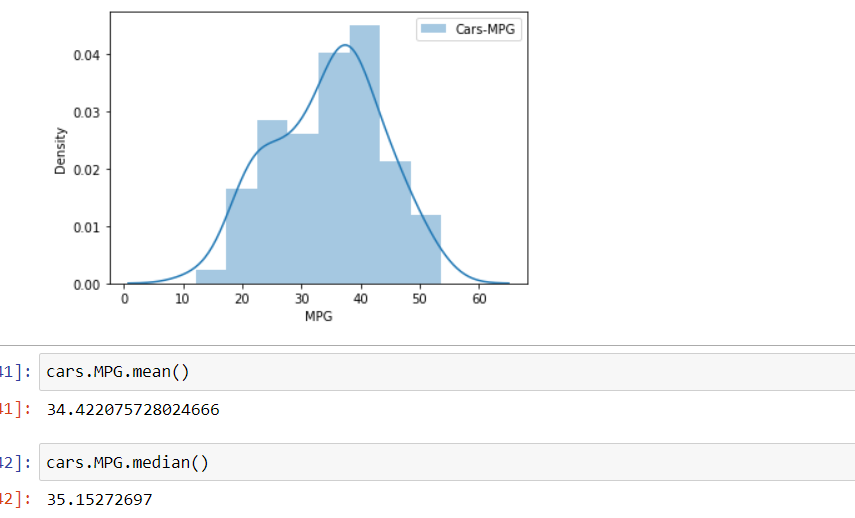
c. P (20<MPG<50)

(round(stats.norm.cdf(50,data.MPG.mean(),data.MPG.std()),3))-(1- round(stats.norm.cdf(20,data.MPG.mean(),data.MPG.std()),3)) =0.0130

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

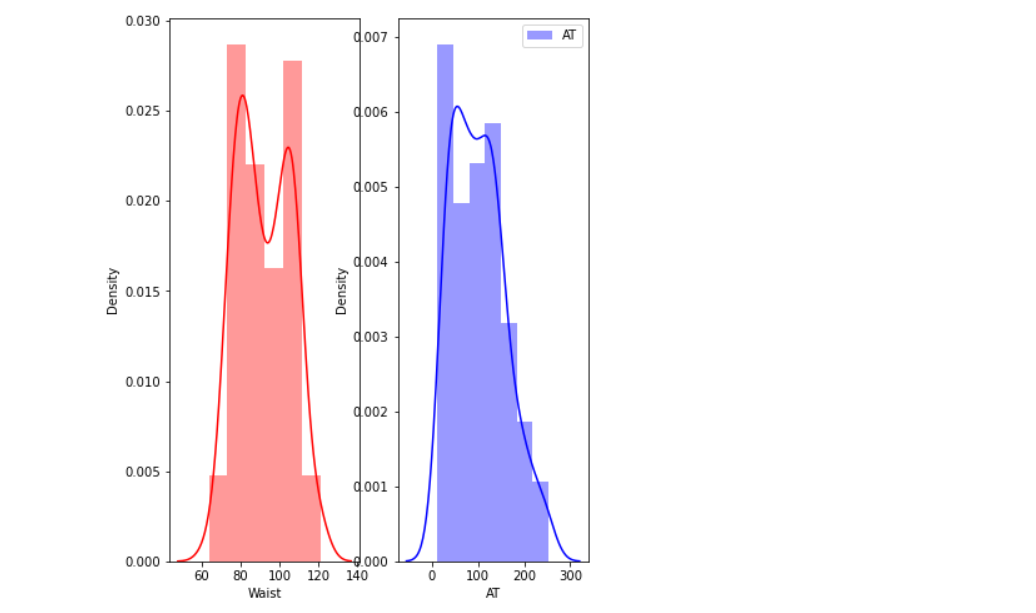
Dataset: Cars.csv



Data of Cars[‘MPG’] is nearly normal distributed , mean and median are close to each other.

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

Dataset: wc-at.csv



print(wcat.Waist.mean()) =92.11

print(wcat.Waist.median())=91.2

print(wcat.AT.mean()) =102.9

print(wcat.AT.median())=96.54

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

Ans:

But the z-curve is designed with only a left tail. So a 90% confidence interval will use the same z-score as 95% of the data.

stats.norm.ppf(0.95) = 1.644

stats.norm.ppf(0.97) =1.8807

stats.norm.ppf(0.8) =0.841

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

# t scores of 95% confidence interval for sample size of 25

stats.t.ppf(0.975,24) # df = n-1 = 24

=2.063

# t scores of 96% confidence interval for sample size of 25

stats.t.ppf(0.98,24)

=2.171

# t scores of 96% confidence interval for sample size of 25

stats.t.ppf(0.98,24)

=2.796

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

solu:

Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

Alternate Hypothesis is: H1= Avg life of Bulb < 260 days

a=(260-270)/(90/np.sqrt(18))

a=-0.471

stats.t.cdf(-0.471,17)

p value =0.3217

Probability of 18 randomly selected bulbs have an average life of no more than 260 days is = 32.17%

Assuming significance value α = 0.05 (Standard Value)(If p\_value > α ; Accept Ho and reject H1)

p-value > α ; Accept Ho i.e. The CEO claims are false and the avg life of bulb > 260 days