|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Numeric discrete data |
| Results of rolling a dice | Numeric discrete data |
| Weight of a person | Numeric continuous data |
| Weight of Gold | Numeric continuous data |
| Distance between two places | Numeric continuous data |
| Length of a leaf | Numeric continuous data |
| Dog's weight | Numeric continuous data |
| Blue Color | Categorical data |
| Number of kids | Numeric discrete data |
| Number of tickets in Indian railways | Numeric discrete data |
| Number of times married | Numeric discrete data |
| Gender (Male or Female) | Categorical 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 |
| High School Class Ranking | Ordinal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ratio |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| 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 | Interval |

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

**Ans.**

When three coins are tossed,

Sample space (s) = (HHH, HHT, HTH, HTT, THH, THT, TTH, TTT)

n(s) = 8

When two heads and one tail are obtained,

Sample space (t) = (HHT, HTH, THH)

n(t) = 3

Probability of obtaining two heads and one tail = n(t)/n(s) = 3/8 = 0.375

**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**

**Ans.**

Total sample space,

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)]

n(S) = 36

1. **Probability that sum is equal to 1**

A = []

n(A) = 0

P(A) = n(A)/n(S) = 0/36 = 0

1. **Probability that sum is less than or equal to 4**

B = [(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)]

n(B) = 6

P(B) = n(B)/n(S) = 6/36 = 1/6 = 0.166..

1. **Probability that sum is divisible by 2 and 3**

C = [(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)]

n(C) = 6

P(C) = n(C)/n(S) = 6/36 = 1/6 = 0.166..

**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?**

**Ans.**

number of ways of drawing 2 balls out of 7 n(S)=7C2=21

number of ways of drawing 2 balls out of 5 n(S)=5C2=10

Probability=Favourable outcomes/Total number of outcomes

Probability =10/21 =0.476

**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

**Ans:**

Expected values of candies for randomly selected child

= 1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.09

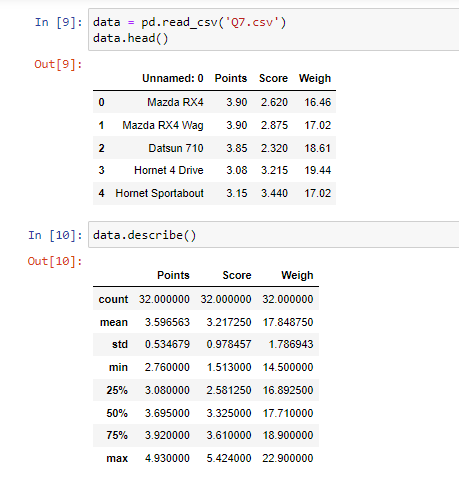
**Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset**

* **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**

**Ans.**

****

**Mean** **for Points** = 3.596

**Mean for Score** = 3.217

**Mean for Weigh** = 17.848

Comment – Mean is the average of some records of numeric data. We calculate mean by adding all numeric data and dividing it by the number of records given.

**Median for Points** = 3.695

**Median for Score** = 3.325

**Median for Weigh** = 17.710

Comment – When there is outlier in our data, the mean value isn’t trustable because difference between mean with outlier and mean without outlier is large. In this case, we calculate median by arranging the numeric data in ascending order and then finding the middle value of all data when there is odd number of records and finding the average of two middle values of data when there is even number of records.

**Mode for Points** = 3.07, 3.92

**Mode for Score** = 3.44

**Mode for Weigh** = 17.02, 18.9

Comment – Mode is most frequently occurring value in our data set. Mode can be multiple values.

**Standard Deviation for Points** = 0.534679

**Standard Deviation for Score** = 0.978457

**Standard Deviation for Weigh** = 1.786943

Comment – Standard Deviation is calculated by taking square root of the Variance of data set. Standard Deviation says how far is each data point from the mean value.

**Variance for Points** = (0.534679)2 = 0.286

**Variance for Score** = (0.978457)2 = 0.957

**Variance for Weigh** = (1.786943)2 = 3.193

Comment – Variance can be calculated by substituting mean from each data point and taking square of each substituted value and then finding the average of all squared values.

**Range for Points** = 4.93 – 2.76 = 2.17

**Range for Score** = 5.424 – 1.513 = 3.911

**Range for Weigh** = 22.90 – 14.50 = 8.4

Comment – Range is the difference between Maximum value and Minimum value of a data set. Range is calculated by substituting the minimum value of all data point from the maximum value of all data point.

**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.**

Average = (108+110+123+134+135+145+167+187+199)/9 = 1308/9 = 145.33

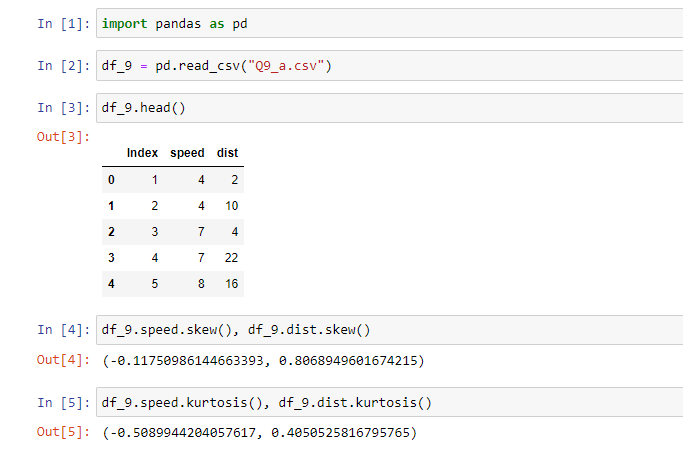
Expected Value of the Weight of that patient = 145.33

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

**Car’s speed and distance**

**Use Q9\_a.csv**

**Ans.**

****

**Skewness:**

Car Speed = - 0.1175

Distance = 0.8069

It defines symmetry of the distribution. If there is outlier in data set, the distribution graph will be asymmetric (not in bell shaped) and large value will be there (in x-axis) in left side or right side of the graph.

**Kurtosis:**

Car Speed = - 0.509

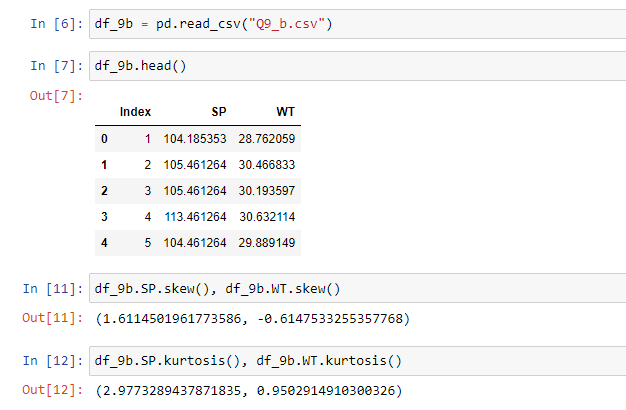
Distance = 0.405

Kurtosis defines the peaked ness of the distribution. Positive kurtosis tells us that distribution is peaked with skinny tails having more outliers. Negative kurtosis tells us that distribution is flat with flat tails having less outliers.

**SP and Weight (WT)**

**Use Q9\_b.csv**

**Ans.**

****

**Skewness:**

SP = 1.61145

WT = - 0.61475

**Kurtosis:**

SP = 2.9773

WT = 0.950

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



**Ans.** From Histogram**:** It isPositively Skewed (Right Skewed) distribution. There is definite probability of having extremely high values data in the dataset as the distribution graph is asymmetric and too narrow in the right side.

From Boxplot**:** The distribution is asymmetric. The number of outliers with extremely high values in the data set is 7 as there is 7 dots in the boxplot.

**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.** Sample size = 2000

Sample mean = 200 pounds

Sample standard deviation = 30 pounds

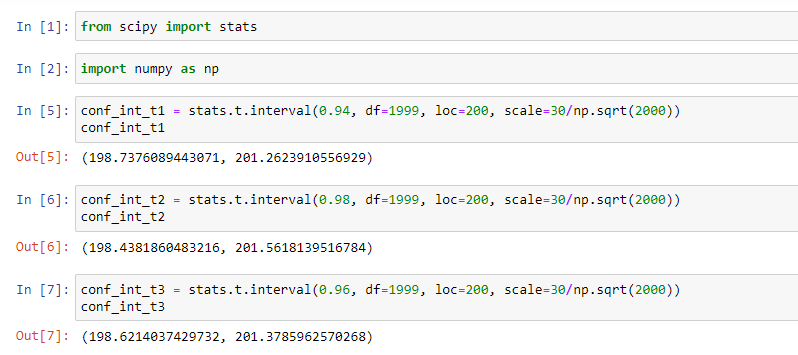
Degree of freedom = 1999

In this case, we use t-test as population standard deviation is not given

94% confidence interval = (198.7376089443071, 201.2623910556929)

98% confidence interval = (198.4381860483216, 201.5618139516784)

96% confidence interval = (198.6214037429732, 201.3785962570268)



**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?

**Ans: 1)**

Mean = (34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56)/18

= 738/18 = 41

Median = (40+41)/2 = 40.5

Variance =

[(34-41)2 + (36-41)2 + (36-41)2 + (38-41)2 + (38-41)2 + (39-41)2 + (39-41)2 + (40-41)2 + (40-41)2 + (41-41)2 + (41-41)2 + (41-41)2 + (41-41)2 + (42-41)2 + (42-41)2 +(45-41)2 + (49-41)2 +(56-41)2] / 18

= [49+25+25+9+9+4+4+1+1+0+1+1+16+84+225] / 18

= 454/18 = 25.22

Standard deviation = sqrt of 25.22 = 5.02

**2)** We can say about the student that scores obtained by the student is normally distributed and each data points are on and average 4.91 marks away from average score.

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

**Ans.** When mean, median of data are equal, then the data is normally distributed and there is no skewness in that data.

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

**Ans.** When mean > median, the nature of skewness is positive skewed also called right skewed.

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

**Ans.** When median > mean, the nature of skewness is negatively skewed also called left skewed.

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

**Ans.** Positive kurtosis value indicates that the distribution is peaked also called Leptokurtic. Leptokurtic has very long and skinny tails, which means there is more chances of having outliers.

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

**Ans.** Negative kurtosis value indicates that the distribution is flatter (less peaked) also called Platykurtic. Platykurtic reveals a distribution with flat tails indicates the small outliers in a distribution.

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



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

**Ans.** The distribution is asymmetric as the median is not in the middle of IQR. Distance between lower quartile and lower extreme is more than distance between upper quartile and upper extreme.

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

**Ans.** Nature of skewness of the data is Negatively skewed or left skewed.

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

**Ans.** Lower quartile, Q1 = 10

Upper quartile, Q3 = 18

Interquartile range, IQR = Q3 – Q1 = 8

**Q19) Comment on the below Boxplot visualizations?**



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

**Ans.** In Boxplot 1, the range of data is smaller (the distance between minimum value and maximum value is small) and there is no outlier in the data. The Interquartile range is small and majority of data are around median. Data are normally distributed and there is no skewness as median (=262.5) is in the middle of the box and range of both whisker is same.

In Boxplot2, the range of data is larger (the distance between minimum value and maximum value is large) and there is no outlier in the data. The Interquartile range is big in that data. Data are normally distributed and there is no skewness as median (=262.5) is in middle of the box and range of both whisker is same.

**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.**

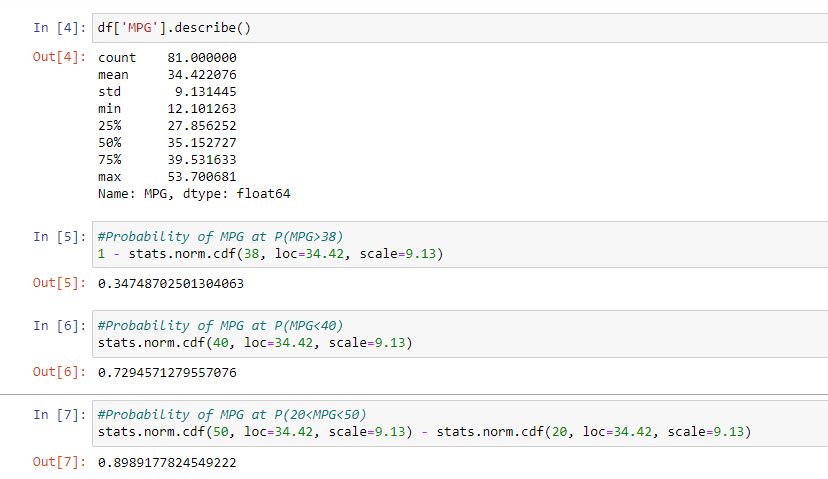
**MPG <- Cars$ MPG**

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

**c. P(20<MPG<50)**

**Ans.**

1. P(MPG>38) = 0.3475
2. P(MPG<40) = 0.7294
3. P(20<MPG<50) = 0.8989



**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

**Ans.** From the histogram, it can be seen that the graph is normally distributed. So, the data follows normal distribution.

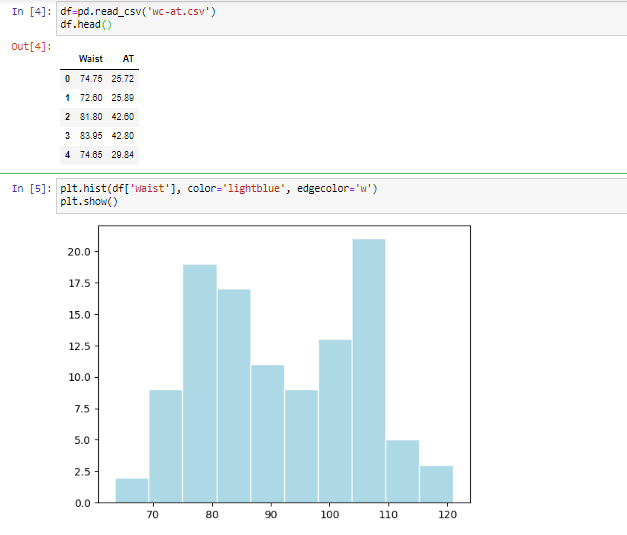


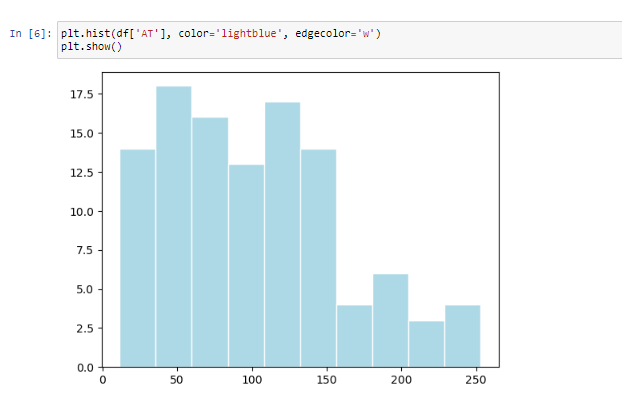
1. **Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution**

**Dataset: wc-at.csv**

**Ans.**

Plotting Histogram for Waist Circumference (Waist): From the histogram, it can be seen that Waist Circumference doesn’t follow Normal Distribution.



Plotting Histogram for Adipose Tissue (AT): 

From the histogram, it can be seen that Adipose Tissue doesn’t follow Normal Distribution.

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

**Ans.**

Using Z table:

Z score value for 90% confidence interval is +- 1.64 (alpha/2 = 0.05)

Z score value for 94% confidence interval is +- 1.88 (alpha/2 = 0.03)

Z score value for 60% confidence interval is +- 2.05 (alpha/2 = 0.20)

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

**Ans.**

Using T table:

T score value for 95% confidence interval with df=24 is +- 2.064

T score value for 96% confidence interval with df=24 is +- 2.057

T score value for 99% confidence interval with df=24 is +- 2.797

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

**Ans.**

Population mean = 270

Sample size =18

Sample mean = 260

Sample standard deviation = 90

As population standard deviation is not given, we use t-distribution to find the probability.

t=(Sample mean-Population mean)/(Sample standard deviation/sqrt(n))

t=(260-270)/(90/sqrt(18))

t=-0.47

Degree of freedom (df) =Sample size-1

Degree of freedom (df) =18-1

Degree of freedom (df) =17

Left tailed test t=-0.424 df=17

From normal distribution table

P\_value=0.3372