|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data |
| Results of rolling a dice | Discrete data |
| Weight of a person | Continuous data |
| Weight of Gold | Continuous data |
| Distance between two places | Continuous data |
| Length of a leaf | Continuous data |
| Dog's weight | Continuous data |
| Blue Color | Discrete data |
| Number of kids | Discrete data |
| Number of tickets in Indian railways | Discrete data |
| Number of times married | Discrete data |
| Gender (Male or Female) | Discrete 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 | 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-Ordinal |
| Level of Agreement | Discrete-Interval |
| IQ(Intelligence Scale) | Discrete-Interval |
| Sales Figures | Discrete-Interval |
| Blood Group | Discrete-Ratio |
| Time Of Day | Continuous-Interval |
| Time on a Clock with Hands | Continuous-Interval |
| Number of Children | Discrete-Interval |
| Religious Preference | Discrete-Ratio |
| Barometer Pressure | Continuous-Interval |
| SAT Scores | Continuous-Ratio |
| Years of Education | Discrete-Nominal |

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

**Answer :**

P(H H T)+P(H T H)+P(T H H)

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

=3/8

Or (Alternative approach)Binomial Distribution.

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 2and 3.

**Answer :**

a) :- Favorable outcomes (sum equal to 1)=0

i.e. not possible that sum always exceed to 1

Required Probability = 0/36=0

b) :- (1,3) (2,2) (3,1) = 3 outcomes, 3/36 i.e. 1/12

C :- Favorable outcomes = sum of divisible by 2 and 3

Sum should be divisible by both 2 and 3

Favorable outcomes = (1,5),(2,4),(3,3),(4,2),(5,1),(6,6)

Therefore , Number of favorable outcomes = 6

i.e. Required Probability = 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 :**

P(2R, 3G, 2B)

P(5/7,4/6) = 20/42

i.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 :**

Expected number of candies for randomly selected child =

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

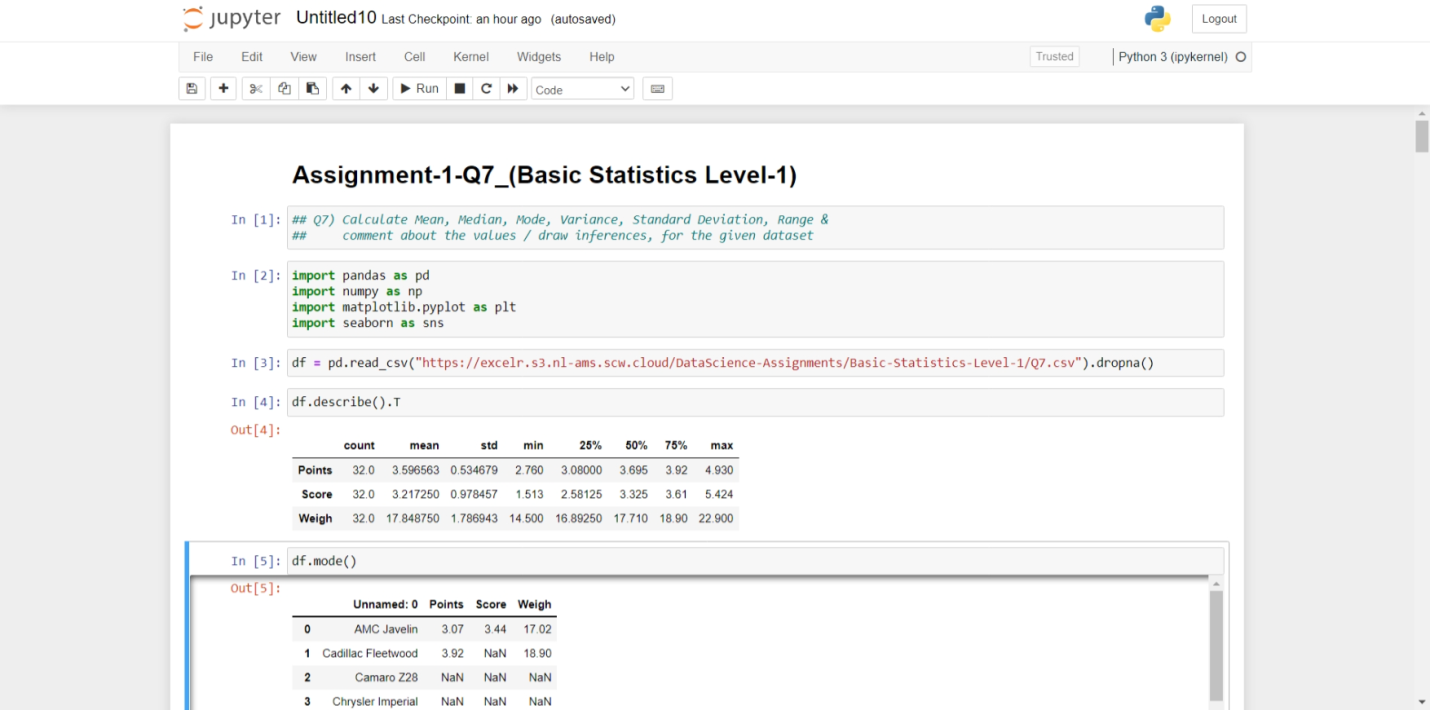
Points:Mean = 3.596563,Median=3.695,Mode=”NaN”, Variance=0.2858814,

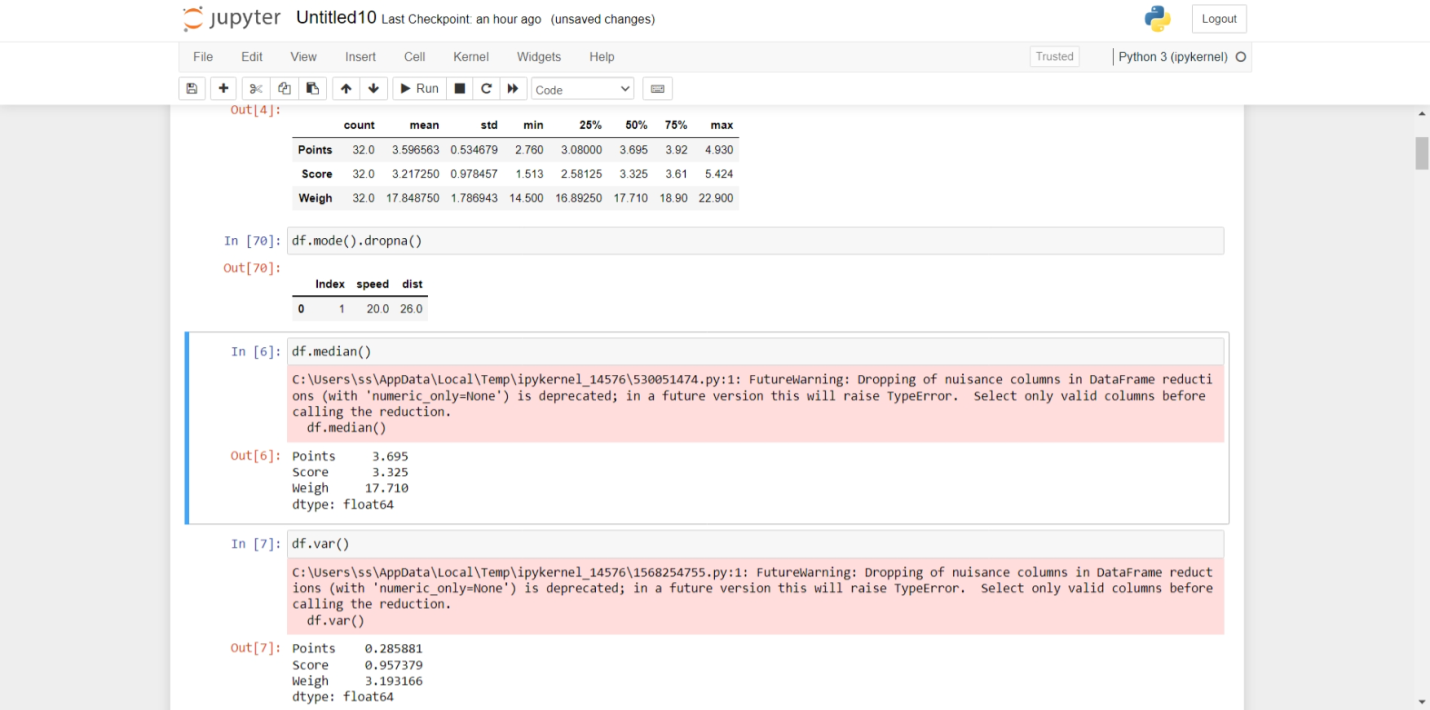
Standard deviation=0.5346787.

Score: Mean = 3.21725, Median = 3.325, Mode = “NaN”, Variance = 0.957379

Standard deviation = 0.9784574

Weight : Mean = 17.84875, Median = 17.71, Mode = “NaN”, Variance = 3.193166, Standard Deviation = 1.786943



****

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 =

Probability of selecting each patient(x) = 1/9

E(x) = {108,110,123,134,135,145,167,187,199}

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**

**Answer :**

A)  Index, Speed and Distance

Skewness of Index is: 0.0

Skewness of Speed is: -0.11395477012828319

Skewness of Distance is: 0.7824835173114966

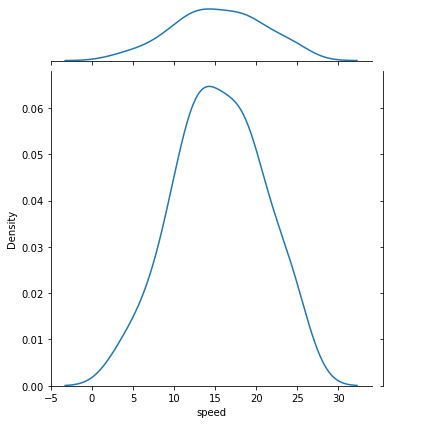
Kurtosis of Index is: -1.2009603841536614

Kurtosis of Speed is: -0.5771474239437371

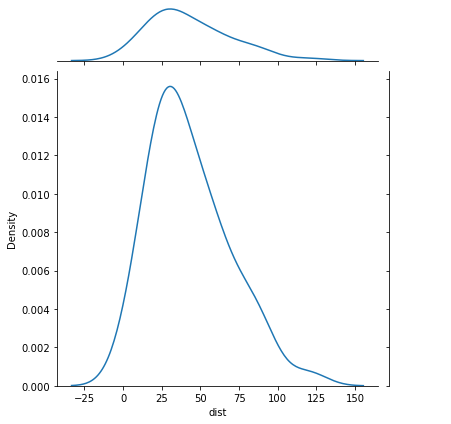
Kurtosis of Distance is: 0.24801865717051808

**Inference**:

The given data of speed are shifted towards right and its tail on the left side, it is a negatively/left-skewed distribution.



The given data of dist are shifted towards left and its tail on the right side, it is a positively/right-skewed distribution



1. SP and Weight(WT)

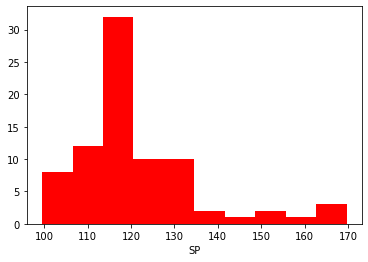
Skewness of SP is: 1.5814536794423764

Skewness of Weight is: -0.6033099322115126

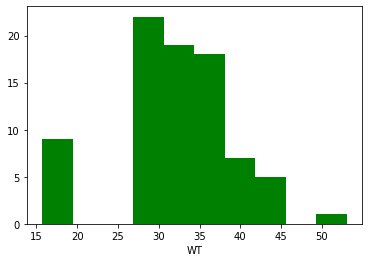
Kurtosis of SP is: 2.7235214865269244

Kurtosis of Weight is: 0.8194658792266849

**Inference:**



Most data points of SP are shifted towards left and its tail on the right side, it is a positively/right-skewed distribution.



Most data points of WT are shifted towards right and its tail on the left side, it is a negatively/left-skewed distribution.

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





**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 :**

Confidence Interval for 94% is :(198.738325292158,201.261674707842)

Confidence Interval for 98% is :(198.439438404299,201.560561595700)

Confidence Interval for96% is :(198.622303348133,201.377696651866)

**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?

**Answer :**

1. Mean : 34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56

/18

=738/18 = 41.0

Median:

(40+41)/2

=81/2 = 40.5

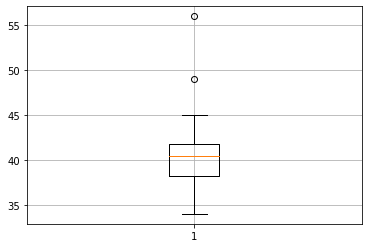
Variance:

(49+25+25+9+9+4+4+1+1+0+0+0+0+1+1+16+64+255)/18

=434/18 = 24.111

Standard Deviation: 4.9103

2)



From above plot we can say that mean of marks of student is 41 which is slightly greater than median. Most of the students got marks in between 40-42, there are two outlier 49, 56.

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

**Answer :**  Symetrical

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

**Answer :**  Right Skewed

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

**Answer :** Left Skewed

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

**Answer :** The data is normally distributed and kurtosis value is 0

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

**Answer :** The distribution of the data has lighter tails and a flatter peaks 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 peak of the given boxplot or the most of data points are situated towards right side and the tail is at left side of boxplot.

What is nature of skewness of the data?

**Answer :**

Negative Skewness

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

**Answer :**

IQR = upper quartile – lower quartile

= 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.

**Answer :**

* By observing both the plots whisker’s level is high in boxplot 2, mean and median are equal hence distribution is symetrical.
* The IQR of the first boxplot is between 255 to 280
* The IQR of the second boxplot is between 225 to 315
* And there is no outliers in both the boxplots

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

Data \_set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

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

c. P (20<MPG<50)

**Answer :**

Probability for MPG>38 is: 0.3475939251582705

Probability for MPG<40 is: 0.7293498762151616

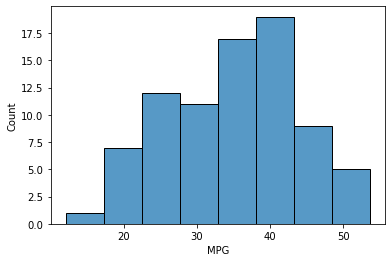
Probability for 20<MPG<50 is: -0.013116469610523339

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Answer :**

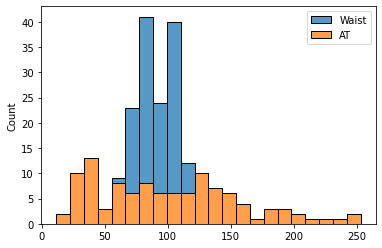


The data set is not a normal distribution,it is right skewed/positively skewed

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

Dataset: wc-at.csv

**Answer :**



Both the Adipose Tissue (AT) and Waist Circumference(Waist) data set do not follow the normal distribution approximately (as mean and median of both the data are approximately different)

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

**Answer :**

90 % confidence interval Z-score: 1.64

94% confidence interval Z-score: 1.55

60% confidence interval Z-score: 0.25

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 t-score:  2.06

96% confidence interval t-score:  2.17

99% confidence interval t-score:  2.8

Q 24**)**A Government companyclaims 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

**Answer :**

T\_Score is-0.4714045207910317

P\_Score is-0.32167411684460556