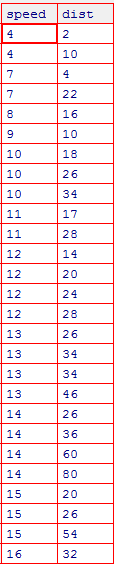
**Name: - UPENDRA DAMA**

**Batch: - 25/01/2020 (Weekend)**

**Module: - 3**

Q1) Calculate Skewness, Kurtosis & draw inferences on the following data

a. Cars speed and distance

****

**Explanation: -**

***Skewness -*** is a measure of the asymmetry of the probability distribution of a random variable about its mean.

If skewness is less than -1 or greater than 1, the distribution is highly skewed.

If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.

If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

In this case, for speed the skewness is -0.1139548 which means the data distribution is skewed left. So, slight magnitude of the data distribution is concentrated on right side and left tail is longer. We can observe same in the below histogram plot.



For distance the skewness is 0.7824835 which means the data distribution is skewed right. So, slight magnitude of the data distribution is concentrated on left side and right tail is longer. We can observe same in the below histogram plot.



***Kurtosis -*** Kurtosis tells you the height and sharpness of the central peak, relative to that of a standard bell curve.Kurtosis is 3 or normal distribution.

If the kurtosis is greater than 3, then the dataset has heavier tails than a normal distribution (more in the tails).

If the kurtosis is less than 3, then the dataset has lighter tails than a normal distribution (less in the tails).

In this case, for speed the Kurtosis is 2.422853 which is less than 3 that means tails are shorter and thinner, and often its central peak is lower and broader. We can see same in the below density plot.

Since Python follows excess kurtosis which is (kurtosis – 3), the kurtosis value in this case is -0.5089944204057617 which is negative kurtosis that means wider peak and thinner tails as shown below.



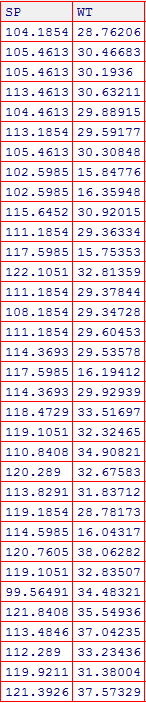
For distance the Kurtosis is 3.248019 which is greater than 3 that means tails are longer and fatter, and often its central peak is higher and sharper. We can see same in the below density plot.

Since Python follows excess kurtosis which is (kurtosis – 3), the kurtosis value in this case is 0.4050525816795765 which is positive kurtosis that means sharp peak and longer tails as shown below.



|  |  |
| --- | --- |
| **R Code** | **Python Code** |
| > library(readr)  > Q1\_a <- read\_csv("Desktop/Digi 360/Module 3/DataSets/Q1\_a.csv")  Parsed with column specification:  cols(  Index = col\_double(),  speed = col\_double(),  dist = col\_double()  )  > View(Q1\_a)  > attach(Q1\_a)  >  > attach(Q1\_a)  The following objects are masked from Q1\_a (pos = 3):  dist, Index, speed  > install.packages("moments")  trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/moments\_0.14.tgz'  Content type 'application/x-gzip' length 55460 bytes (54 KB)  ==================================================  downloaded 54 KB  The downloaded binary packages are in  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T//RtmpLTvveY/downloaded\_packages  >  > library(moments)  > #Find skewness for Speed  > skewness(speed)  [1] -0.1139548  > #Find skewness for Distance  > skewness(dist)  [1] 0.7824835  > #Find Kurtosis for speed  > kurtosis(speed)  [1] 2.422853  > #Find Kurtosis for distance  > kurtosis(dist)  [1] 3.248019  > #grphical representation  > hist(speed)    > hist(dist)    > install.packages("UsingR")  > densityplot(speed)    > densityplot(dist) | import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  import seaborn as sns  assignment\_md3 = pd.read\_csv ("~/desktop/Digi 360/Module 3/DataSets/Q1\_a.csv")  print("Skewness for Speed:", assignment\_md3.speed.skew())  print("Skewness for Dist:", assignment\_md3.dist.skew())  print("Kurtosis for Speed:", assignment\_md3.speed.kurt())  print("Kurtosis for dist:", assignment\_md3.dist.kurt())  #Graphical Representation  plt.hist(assignment\_md3.speed)  plt.title("Histogram for speed")  plt.show()  plt.hist(assignment\_md3.dist)  plt.title("Histogram for dist")  plt.show()  sns.distplot(assignment\_md3.speed)  plt.title("Density for Speed")  plt.show()  sns.distplot(assignment\_md3.dist)  plt.title("Density for Distance")  plt.show()  **Output: -**  Skewness for Speed: -0.11750986144663393  Skewness for Dist: 0.8068949601674215  Kurtosis for Speed: -0.5089944204057617  Kurtosis for dist: 0.4050525816795765  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T/com.microsoft.Word/Content.MSO/A11264DD.tmp  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T/com.microsoft.Word/Content.MSO/DCD567F3.tmp  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T/com.microsoft.Word/Content.MSO/C7E5979.tmp  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T/com.microsoft.Word/Content.MSO/1921F2EF.tmp |

b. Top Speed (SP) and Weight (WT)

****

**Explanation: -**

***Skewness: -***

In this case, for top speed the skewness is 1.581454 which means the data distribution is skewed right. Since the value is greater than 1, mass of the data distribution is concentrated on left side and right tail is longer. We can observe same in the below histogram plot.



For weight the skewness is -0.6033099 which means the data distribution is skewed left. So, slight magnitude of the data distribution is concentrated on right side and left tail is little longer. We can observe same in the below histogram plot.



***Kurtosis: -***

In this case, for top speed the Kurtosis is 5.723521 which is greater than 3 that means tails are longer and fatter, and often its central peak is higher and sharper. We can see same in the below density plot.

Since Python follows excess kurtosis which is (kurtosis – 3), the kurtosis value in this case is 2.9773289437871764 which is positive kurtosis that means sharp peak and longer tails as shown below.



For distance the Kurtosis is 3.819466 which is also greater than 3 that means tails are longer and fatter, and often its central peak is higher and sharper. We can see same in the below density plot.

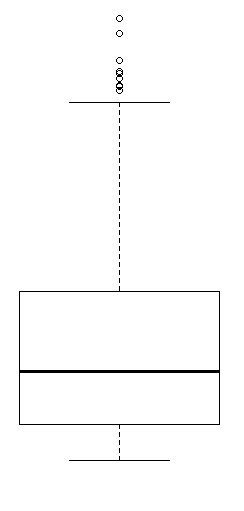
Since Python follows excess kurtosis which is (kurtosis – 3), the kurtosis value in this case is 0.9502914910300326 which is positive kurtosis that means sharp peak and longer tails as shown below.



|  |  |
| --- | --- |
| **R Code** | **Python Code** |
| > library(readr)  > Q2\_b <- read\_csv("Desktop/Digi 360/Module 3/DataSets/Q2\_b.csv")  Parsed with column specification:  cols(  X1 = col\_double(),  SP = col\_double(),  WT = col\_double()  )  > View(Q2\_b)  > attach(Q2\_b)  >  > attach(Q2\_b)  The following objects are masked from Q2\_b (pos = 3):  SP, WT, X1  > install.packages("moments")  trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/moments\_0.14.tgz'  Content type 'application/x-gzip' length 55460 bytes (54 KB)  ==================================================  downloaded 54 KB  The downloaded binary packages are in  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T//RtmpLTvveY/downloaded\_packages  >  > library(moments)  > #Find skewness for top Speed  > skewness(SP)  [1] 1.581454  > #Find skewness for Weight  > skewness(WT)  [1] -0.6033099  > #Find Kurtosis for top speed  > kurtosis(SP)  [1] 5.723521  > #Find Kurtosis for Weight  > kurtosis(WT)  [1] 3.819466  > #grphical representation  > hist(SP)    > hist(WT)    > install.packages("UsingR")  > densityplot(SP)    > densityplot(WT) | import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  import seaborn as sns  assignment\_md3 = pd.read\_csv ("~/desktop/Digi 360/Module 3/DataSets/Q2\_b.csv")  print("Skewness for top Speed:", assignment\_md3.SP.skew())  print("Skewness for Weight:", assignment\_md3.WT.skew())  print("Kurtosis for top Speed:", assignment\_md3.SP.kurt())  print("Kurtosis for Weight:", assignment\_md3.WT.kurt())  #Graphical Representation  plt.hist(assignment\_md3.SP)  plt.title("Histogram for top speed")  plt.show()  plt.hist(assignment\_md3.WT)  plt.title("Histogram for Weight")  plt.show()  sns.distplot(assignment\_md3.SP)  plt.title("Density for top Speed")  plt.show()  sns.distplot(assignment\_md3.WT)  plt.title("Density for Weight")  plt.show()  **Output: -**  Skewness for top Speed: 1.6114501961773555  Skewness for Weight: -0.6147533255357768  Kurtosis for top Speed: 2.9773289437871764  Kurtosis for Weight: 0.9502914910300326  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T/com.microsoft.Word/Content.MSO/6818A9D5.tmp  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T/com.microsoft.Word/Content.MSO/DF020FAB.tmp  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T/com.microsoft.Word/Content.MSO/C07F31F1.tmp  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T/com.microsoft.Word/Content.MSO/5DBBFA27.tmp |

Q2) Draw inferences about the following boxplot & histogram





**Explanation: -**

***Histogram: -*** We can define below points by looking at above histogram.

*Shape: -* By looking at the above histogram we can easily say that the shape is right skewed.Because most of the chick weight is concentrated at the left side.

*Center: - The median of the above histogram is approximately 200 and mean would be approximately between 250 and 300. The mode is 75. So, in this case Mode < Median < Mean which tells it is right skewed.*

*Spread: -* The chicken weight is distributed within the range of 0 to 400. But most of the chick weight concentrated in the range 50-100 with frequency 200.

*Skew: -* The histogram is skewed right which means the tail is going off to the right. Most of the chick weight is away from the center and concentrated at the left side.

*Modality:* - Since we see only one peak, this histogram is unimodal.

*Outliers: -* There are no outliers since we don’t see any isolated bar in the histogram.

***Boxplot: -*** We can define below points by looking at above boxpot.

*Shape: - The long part of the box is above the median, so the distribution is right skewed.*

*Center: - The thick line at the middle of the box is median. Here median is less than mean.*

*Range: - It can be represented by the distance between smallest value and largest value including outliers.*

*Spread: - There are less data points between first quartile Q1 and lower extreme point. More data points between Q3 and upper extreme point. Spread can be measured by IQR which is Q3 – Q1.*

*Outliers: -We see there are outliers above the upper(second) whisker.*

**Q3)** 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?

**Explanation: -**

Parameters given are –

N = 3,000,000

n = 2000

s = 30

= 200

Since we don’t have σ, we need to go with student t distribution formula.

Confidence interval of population mean = t1-α,1-n s/

*94% confidence Interval: -*

*1 - α = 94%*

*Degrees of freedom = n-1 = 1999*

***#t value calculation: -***

*> qt(0.97,1999)*

*[1] 1.881861*

***#t distribution calculation: -***

*> qt(0.97,1999,200,30)*

*[1] 206.4483*

*> qt(0.03,1999,200,30)*

*[1] 193.9561*

Confidence interval of population mean = [206.4483, 193.9561]

*98% confidence Interval: -*

*1 - α = 98%*

*Degrees of freedom = n-1 = 1999*

***#t value calculation: -***

*> qt(0.99,1999)*

*[1] 2.328215*

***#t distribution calculation: -***

*> qt(0.99,1999,200,30)*

*[1] 208.0254*

*> qt(0.01,1999,200,30)*

*[1] 192.567*

Confidence interval of population mean = [208.0254, 192.567]

*96% confidence Interval: -*

*1 - α = 96%*

*Degrees of freedom = n-1 = 1999*

***#t value calculation: -***

*> qt(0.98,1999)*

*[1] 2.05509*

***#t distribution calculation: -***

*> qt(0.98,1999,200,30)*

*[1] 207.0579*

*> qt(0.02,1999,200,30)*

*[1] 193.4148*

Confidence interval of population mean = [207.0579, 193.4148]

**Q4)** 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?

**Explanation: -**

1. Mean = 41, Median = 40.5, Variance = 25.53 and Standard deviation = 5.052.
2. Mean is greater than the median so the shape of the distribution will be skewed towards right. That means majority of marks are concentrated on the left side. Variance is 25.52 that means the student scores 25.52 far spread out from their mean.

|  |  |
| --- | --- |
| **R Code** | **Python Code** |
| > x <- c(34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)  > mean(x)  [1] 41  > median(x)  [1] 40.5  > var(x)  [1] 25.52941  > sd(x)  [1] 5.052664 | import numpy as np  dataset = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]  print("Mean:", np.mean(dataset))  print("Median:", np.median(dataset))  print("Variance:", np.var(dataset))  print("St deviation:", np.std(dataset))  **Output: -**  Mean: 41.0  Median: 40.5  Variance: 24.11111111111111  St deviation: 4.910306620885412 |

Q5) What is the nature of skewness when mean, median of data are equal?

**Answer: -** *If mean is equal to median then skewness will be zero that means the distribution will be symmetrical.*

Q6) What is the nature of skewness when mean > median?

**Answer: -** *If mean is greater than median the distribution will be right skewed.*

Q7) What is the nature of skewness when median > mean?

**Answer: -** *If median is greater than mean the distribution will be left skewed.*

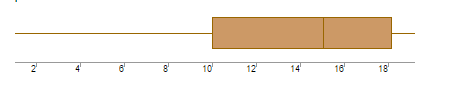
Q8) What does positive kurtosis value indicates for a data?

**Answer: -** *Tails are longer and fatter, and often its central peak is higher and sharper.*

Q9) What does negative kurtosis value indicates for a data?

**Answer: -** *Tails are shorter and thinner, and often its central peak is lower and broader*

Q10) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

***Answer: -*** *Let’s assume above box plot is about hights of the plants (in cm.) in a garden. 50% of the plants are above 10 cm height and remaining are less. Also plants whose height is above 15cm are approx. 40%. The plants heights are distributed within the range of 18 cm ( 19 -1).*

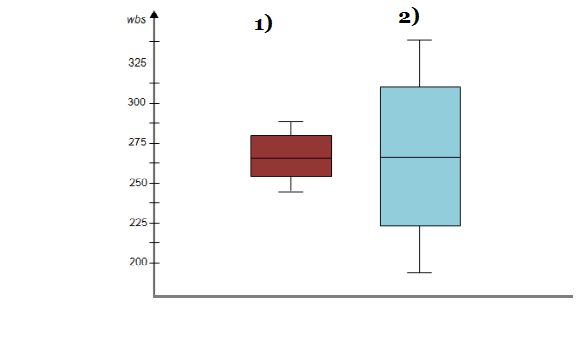
What is nature of skewness of the data?

***Answer: -*** *Here median is greater than mean. So, the distribution is left skewed.*

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

***Answer: -*** *Q3 18 and Q1 10. The IQR = Q3 – Q1 8*

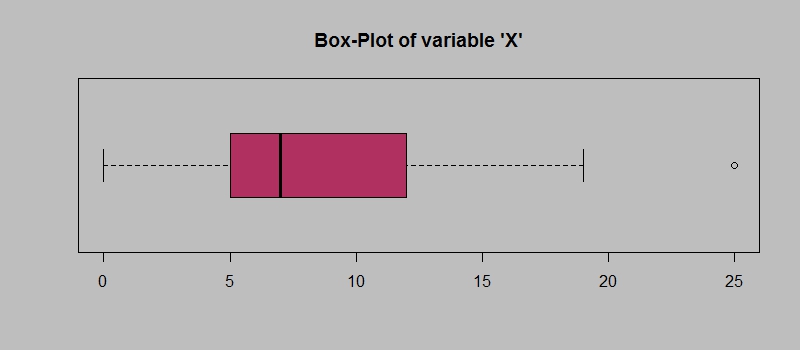
Q11) Comment on the below Boxplot visualizations?



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 symmetrical for both the plots. Median for both the plots are same but the range (50) of the distribution is low for plot 1 when compared with range (150) of plot 2.*

Q12)



Answer the following three questions based on the boxplot above.

1. What is inter-quartile range of this dataset? (please approximate the numbers) In one line, explain what this value implies.

***Answer: -*** *Q3 12 and Q1 5. The IQR = Q3 – Q1 7 which means that the data distributed around from 5 to 12 with median around 7.*

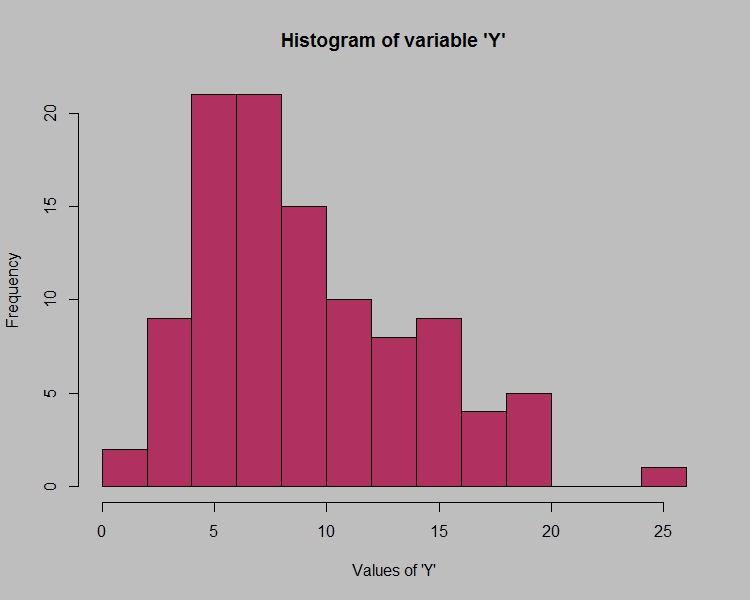
1. What can we say about the skewness of this dataset?

***Answer: -*** *Here mean is greater than median. So, the distribution is right skewed.*

1. If it was found that the data point with the value 25 is actually 2.5, how would the new boxplot be affected?

***Answer: -*** *Q1 position will move to left side to number 2.5 and outlier will not be described into graph. The box will be symmetrically divided.*

Q13)



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

***Answer: -*** *The mode lies in two numbers (5 and 7) which means it is bimodal.*

1. Comment on the skewness of the dataset.

***Answer: -*** *The skewness of the histogram moves towards right side and majority of the data is concentrated on left side.*

1. Suppose that the above histogram and the boxplot in question 2 are plotted for the same dataset. Explain how these graphs complement each other in providing information about any dataset.

***Answer: -*** *The box plot shows the median, q1 and q3 also the outliers. Whereas the histogram explains the mode or frequency with outliers identified.*