**Exploratory Data Analysis**

Instructions:

Please share your answers filled inline in the word document. Submit Python/R code files wherever applicable.

Please ensure you update all the details:

**Name: Upadhyay Sachin Naresh**

**Batch Id: Data Science\_08032021**

**Topic: Exploratory Data Analysis**

**Problem Statements:**

Q1) Calculate Skewness, Kurtosis using R/Python code & draw inferences on the following data.

**Hint:** [Insights drawn from the data such as data is normally distributed/not, outliers, measures like mean, median, mode, variance, std. deviation]

**a. Cars speed and distance**

****

**Answer:**

**Input:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from scipy.stats import kurtosis

from scipy.stats import skew

from scipy import stats

import math

data = pd.read\_csv("C:/Users/usach/Desktop/Statistical Datasets/Q1\_a.csv")

data1=data[['speed','dist']]

data1.describe()

data1.skew()

data1.kurtosis()

a=stats.kurtosis(data1)

b=stats.skew(data1)

plt.plot(a,b)

stats.describe(data1**)**

**Output:**

data1=data[['speed','dist']]

data1.describe()

Out[29]:

speed dist

count 50.000000 50.000000

mean 15.400000 42.980000

std 5.287644 25.769377

min 4.000000 2.000000

25% 12.000000 26.000000

50% 15.000000 36.000000

75% 19.000000 56.000000

max 25.000000 120.000000

[IN]:data1.skew()

Out[30]:

speed -0.117510

dist 0.806895

dtype: float64

data1.kurtosis()

Out[31]:

speed -0.508994

dist 0.405053

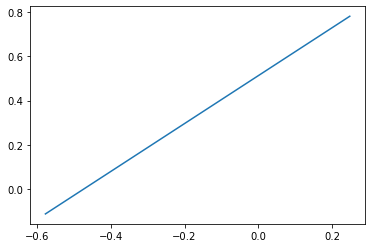
dtype: float64

a=stats.kurtosis(data1)

b=stats.skew(data1)

plt.plot(a,b)

Out[34]: [<matplotlib.lines.Line2D at 0x20e26a67580>]



stats.describe(data1)

Out[36]: DescribeResult(nobs=50, minmax=(array([4, 2], dtype=int64), array([ 25, 120], dtype=int64)), mean=array([15.4 , 42.98]), variance=array([ 27.95918367, 664.06081633]), skewness=array([-0.11395477, 0.78248352]), kurtosis=array([-0.57714742, 0.24801866]))

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

****

**Answer:**

**Input:**

data = pd.read\_csv("C:/Users/usach/Desktop/Statistical Datasets/Q2\_b.csv")

data2=data[['SP','WT']]

data2.kurtosis()

data2.skew()

data2.describe()

stats.describe(data2)

x=stats.kurtosis(data)

y=stats.skew(data)

plt.plot(x,y)

**Output:**

data2.kurtosis()

Out[39]:

SP 2.977329

WT 0.950291

dtype: float64

data2.skew()

Out[40]:

SP 1.611450

WT -0.614753

dtype: float64

data2.describe()

Out[41]:

SP WT

count 81.000000 81.000000

mean 121.540272 32.412577

std 14.181432 7.492813

min 99.564907 15.712859

25% 113.829145 29.591768

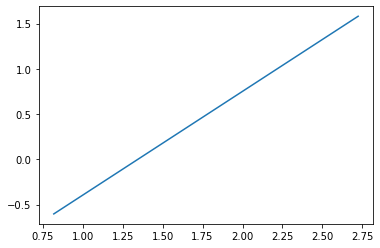
50% 118.208698 32.734518

75% 126.404312 37.392524

max 169.598513 52.997752

stats.describe(data2)

Out[42]: DescribeResult(nobs=81, minmax=(array([99.56490661, 15.71285853]), array([169.5985128 , 52.99775236])), mean=array([121.54027218, 32.41257691]), variance=array([201.1130015 , 56.14224661]), skewness=array([ 1.58145368, -0.60330993]), kurtosis=array([2.72352149, 0.81946588]))



Q2) Draw inferences about the following boxplot & histogram.

**Hint:** [Insights drawn from the plots about the data such as whether data is normally distributed/not, outliers, measures like mean, median, mode, variance, std. deviation]





**Answer:**

* The histogram show the completely values falls on the positive side it means it contains positive skew and there is no outliers in the histogram.
* Box plot contains the outliers and the maximum values are present in the upper quartile. The top whisker is much longer than the bottom whisker and the line is gravitating towards the bottom of the box.

Q3) 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? [**Hint**: Looking at the various measures calculated above whether the data is normal/skewed or if outliers are present].

**Answer:**

**1)**

**Input:**

a = c(34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)

mean(a)

median(a)

var(a)

sd(a)

x=pd.DataFrame(tests).skew()[0]

x

**Output:**

> mean(a)

[1] 41

> median(a)

[1] 40.5

> var(a)

[1] 25.52941

> sd(a)

[1] 5.05266

2)The data skewed and it is positive skewness and yes outliers are present in it.

The skewness value in the marks has 1.68 it means it contains positive skew



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

**Answer:**

Nature of Skewness: Skewness can be positive, negative or zero. When the value of mean, media and mode are equal there is no skewness. Skewness is the measure of the asymmetry of probability distribution real named mean.

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

**Answer:**

If the mean is greater than the [median](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median), the distribution is positively skewed.

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

**Answer:** Negative skewness. It is in the negative direction

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

**Answer:**

Positive values of kurtosis indicate that a distribution is peaked and possess thick tails.

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

**Answer:** Negative Kurtosis indicates Wider peak and thinner tails.

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



What can we say about the distribution of the data?

**Answer:**  Negative skew

What is nature of skewness of the data?

**Answer:** Left Skewed

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

**Answer:** 8  
Q11) Comment on the below Boxplot visualizations?



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

**Hint**: [On comparing both the plots, and check if the data is normally distributed/not, outliers present, skewness etc.]

**Answer:** Both are not Normally Distributed, No outliers, Both are Negative skewed

Q12)



Answer the following three questions based on the boxplot above.

1. What is inter-quartile range of this dataset? [**Hint**: IQR = Q3 – Q1]

In one line, explain what this value implies. (**Hint:** Based on IQR definition)

**Answer:** IQR value is 7 which implies the 50% of the data.

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

**Answer:** Right skewness, Positive.

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

(**Hint:** On changing the data point from 25 to 2.5 in the data, how is it different from the current one.)

**Answer:** 2.5 will be not considered an outlier. The boxplot will start from 0 and send at 20 in representation.

Q13)



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie? **Hint:** [In terms of values On Y-axis]

**Answer:** 4-8

1. Comment on the skewness of the dataset

**Answer:** Positive skewness

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. **Hint:** [Visualizing both the plots, draw the insights]

**Answer:**

Median in boxplot and Mode in histogram Histogram provides the frequency distribution so we can see how many times each data point is occurring however boxplot provides the quantile distributioni.e.50% data lies between 5 and 12. Boxplot provides whisker length to identify outliers, no information from histogram. We can only guess looking at the gap that 25 may be an outlier.

**Hints:**

For each assignment, the solution should be submitted in the below format

1. Research and Perform all possible steps for obtaining solution

2.

3. For Statistics calculations, explanation of the solutions should be documented in black and white along with the codes.

Must follow these guidelines:

3.1. Be thorough with the concepts of Probability, Central Limit Theorem and Perform the

calculation stepwise

3.2. For True/False Questions, or short answer type questions explanation is must

3.3. R & Python code for Univariate Analysis (histogram, box plot, bar plots etc.) the data

distribution to be attached

4. All the codes (executable programs) should execute without errors

5. Code modularization should be followed

6. Each line of code should have comments explaining the logic and why you are using that