Stastics-1 Assignment

Q 1. what is the difference between descriptive stastics and inferential stastics? Explain with examples.

Ans :- Descriptive stastics :- this type deals with describing or summarizing data.

Main tool :-1. Measure of central tendency (mean, median, mode)

2. measure of dispersion (range, variance, standard deviation)

3. graphs and tables (pie chart, bar charts, histogram)

Ex :- a. average marks of a class

b. maximum temperature this week

inferential stastics :-this type helps in making prediction or generalization from a sample to a population.

Main tool :- hypothesis testing

Confidence intervals

Regression, probability distribution

Ex :- prediction result from exit polls

Estimate average salary of engineers in India by surveying 1000 people.

Q2. What is sampling in stastics? Explain the difference between random and stratified sampling.

Ans:-sampling is the process of selecting a smaller group(sample) from a larger population to draw conclusions about the whole group.

Population=entire group

Sample=part of the group we actually study

Why do we use sampling:- faster than studying the whole population cheaper and more resource efficiency practically possible in large scale study.

Random sampling:- every individual Ashan equal chance pf being selected like drawing names from a

Eg:- randomly select100 employee for a company of 1000

```
Import random

Data=list(range(1,1001))

Sample=random. Sample(dtata,10)

Print(“simple random sample:”, sample)
```

Stratified sampling:- divided population into subgroup than randomly sample from each ensure representation of all categories.

Eg:- sample 50 student from each department in college

Real use:- used in survey by age group, gender ,income bracket.

Q 3. Define mean, median and mode. Explain why these measure of central tendency are imported.

Ans :- mean(average) what it is –

Add all the numbers divided by how many there are.

Formula:- mean= x1+x2+x3+...........+xn /n

Eg:- data [5,10,15,20,25] = 5+10+15+20+25/5=15 mean =15

Use:- best used when data is symmetric

With no extreme outliers

Eg:- average marks of a students

Average rainfall over a week

2. median:- the middle value when all values are sorted. If even number of values average of two middle numbers.

Eg:- data [4,6,8,10,2] = 2,4,6,8- 4+6/2 =5

Use:- median income of a city

Median house price in neighbourhood.

3.mode:- the most frequently occurring value in the dataset. A dataset on have no mode,one mode,two mode, more than two mode

Eg:- data [2,4,4,6,8,8,10] mode =8

Use:- most common soe size sold in a store

Most popular pizza topping

Q 4. Explain skewness and kurtosis. What does a positive skew imply about the data?

Ans:- skewness measure the asymmetry of a dataset around its mean. It tells us whether the tail of the data is longer on the left or right side of the distribution.

Why is skewness important?

Help us decide which measure of central tendency is best (mean, median or mode) guides us in choosing the right stastics test. Affects assumption in machine learning models and finance forecasting helps identify the presence of outliers or extreme values

How is skewness measured?

There are many formulas but

Skewness=mean-median/standard deviation

O=perfectly symmetric

>o=right skewed

<=left skewed

Kurtosis:- kurtosis is a stastics measure that describe the shape of tails of a probability distribution compared to a normal distribution. tailenders of a probability distribution, specifically how heavily its tails differ from those of a normal distribution. There are three main types of kurtosis

Mesokurtic:- which has a kurtosis similar to a normal distribution(baseline)

Leptokurtic:- which has heavier tails and more outliers than a normal distribution

Platykurtic:- which has lighter, shorter tails and fewer extreme values.

A positively skewed distribution means the data's tail is longer on the right side, indicating that most values are concentred on the left , with a few high-value outliers pulling the mean to the right. This result in the mean being greater than the median (mean>median)

Q 5. Implement a python program to compute the mean, median and mode of a given list numbers.

data=[12,15,12,8,19,12,20,22,19,24,24,26,28]

```
import NumPy as np
import pandas as pd
import statistics as stats
data=[12,15,12,8,19,12,20,22,19,24,24,26,28]
# calculate mean
mean_value=stats.mean(data)
print("mean using statistics:",mean_value)
mean_np=np.mean(data)
print("mean using numpy:",mean_np)
df=pd.Series(data)
print("mean using pandas:",df.mean())
# calculate median
print("median using stastics:",stats.median(data))
print("median using numpy:",np.median(data))
print("median using pandas:",df.median())
# calculate mode
print("mode using statistics:",stats.mode(data))
print("mode using pandas:",df.mode())
print("mode using pandas:",df.mode()[0])
```

```
Output: mean using statistics: 18.53846153846154
mean using numpy: 18.53846153846154
mean using pandas: 18.53846153846154
median using stastics: 19
median using numpy: 19.0
median using pandas: 19.0
mode using statistics: 12
mode using pandas: 0    12
dtype: int64
mode using pandas: 12
```

Q6.compute the covariance and correlation coefficient between the following two datas provides a s list in python:

```
import pandas as pd

import numpy as np

data=({

“x”:list_x,

“y”:list_y

})  # convert to dataframe

Print(DataFrame)

data={

   'list_x':[10,20,30,40,50],

   'list_y':[15,25,35,45,60]

}

df=pd.DataFrame(data)

# calculate covariance

covarince=df.cov()

print("covariance matrix:\n",covarince)
```

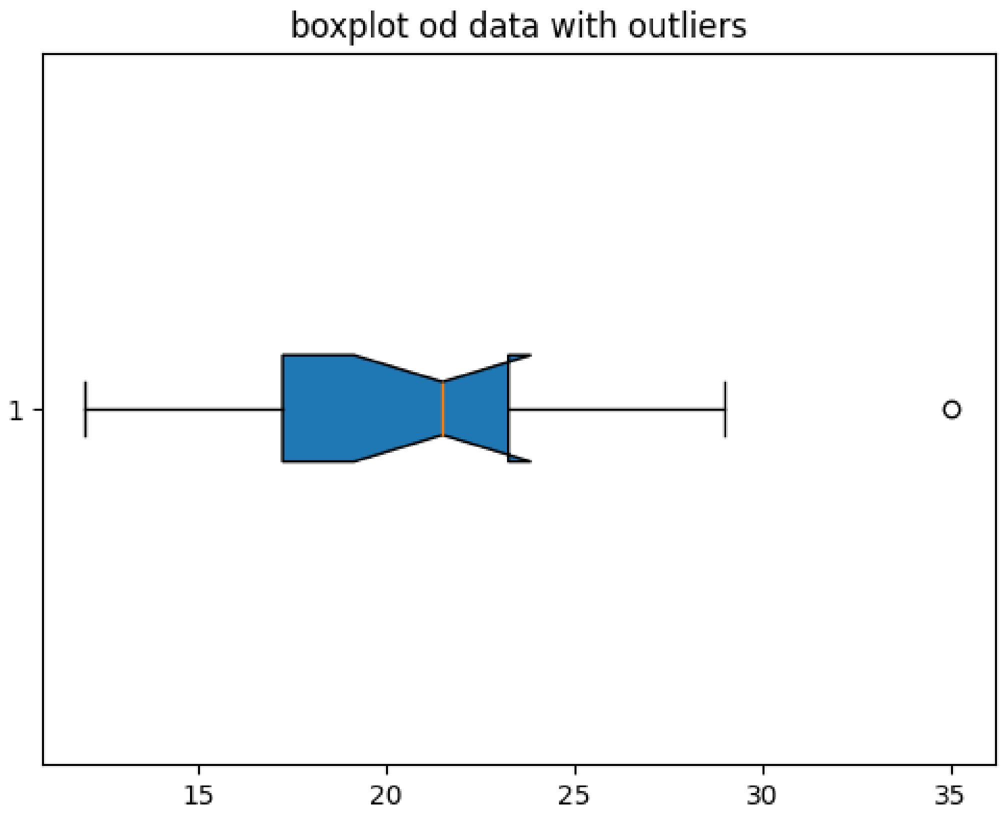
```
Output:-

covariance matrix:

      list_x  list_y

list_x  250.0   275.0

list_y  275.0   305.0

correlation matrix:

        list_x   list_y

list_x  1.000000  0.995893

list_y  0.995893  1.000000
```

Q7

```
import matplotlib.pyplot as plt
import numpy as np
data=[12,14,14,15,18,19,19,21,22,22,23,23,24,26,29,35]
# boxplot
plt.boxplot(data,vert=False,patch_artist=True,notch=True)
plt.title("boxplot od data with outliers")
plt.xlabel('values')
plt.show()
# outliers
Q1=np.percentile(data,25)
Q3=np.percentile(data,75)
IQR=Q3-Q1
lower_bound=Q1-1.5*IQR
upper_bound=Q3+1.5*IQR
outliers=[X for X in data if X<lower_bound or X> upper_bound]
print("Q1:",Q1)
print("Q3:",Q3)
print("IQR:",IQR)
print("lower_bound:",lower_bound)
print("upper-bound:",upper_bound)
print("outliers:",outliers)
```

Output-

```
Q1: 17.25
Q3: 23.25
IQR: 6.0
lower_bound: 8.25
upper-bound: 32.25
outliers: [35]
```

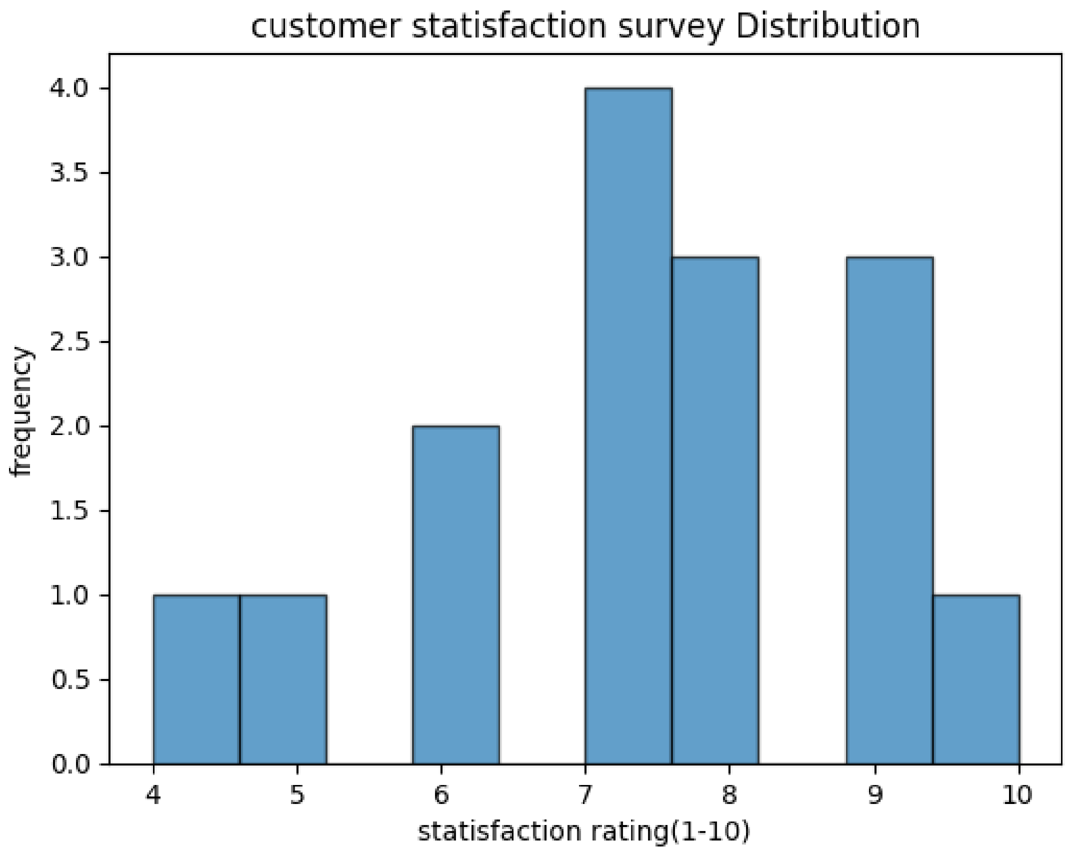


Q8.you are working as a data analyst in an e-commerce company. The marketing team wants to know if there is a relationship between advertising spend and daily sales. Explain how you would use relationship between

```
import pandas as pd

df=pd.DataFrame({

  'advertising_spend':[200,250,300,400,500],

  'daily_sales':[2200,2450,2750,3200,4000]

})

print(":Data")

print(df)

# calculate covariance

cov_matrix=df.cov()

print("covariance matrix:\n",cov_matrix)

# calculate correlation

corr_matrix=df.corr()

print("correlation matrix:\n",corr_matrix)
```

```
Output:-

:Data

  advertising_spend  daily_sales

0           200       2200

1           250       2450

2           300       2750

3           400       3200

4           500       4000

covariance matrix:

              advertising_spend  daily_sales

advertising_spend         14500.0     84875.0

daily_sales              84875.0    503250.0

correlation matrix:

              advertising_spend  daily_sales

advertising_spend        1.000000    0.993582

daily_sales             0.993582    1.000000
```

```
Q9. import matplotlib.pyplot as plt
import numpy as np
import statistics as stats
survey_data=[7,8,5,9,6,7,8,9,10,4,7,6,9,8,7]
# summary statistics
mean_val=np.mean(survey_data)
median_val=np.median(survey_data)
mode_val=stats.mode(survey_data)
std_dev=np.std(survey_data,ddof=1)
min_val,max_val=np.min(survey_data),np.max(survey_data)
q1,q3=np.percentile(survey_data,[25,75])
print("mean:",mean_val)
print("median:",median_val)
print("mode:",mode_val)
print("standard deviation:",std_dev)
print("min:",min_val)
print("max:",max_val)
print("Q1:",q1,"Q3:",q3)


# histogram plot
plt.hist(survey_data,bins=10,edgecolor="black",alpha=0.7)
plt.title("customer statisfaction survey Distribution")
plt.xlabel("statisfaction rating(1-10)")
plt.ylabel("frequency")
plt.grid(axis="y",linestyle="_ _",alpha=0.6)
plt.show()
```



```
Output :-

mean: 7.333333333333333

median: 7.0

mode: 7

standard deviation: 1.632993161855452

min: 4

max: 10

Q1: 6.5 Q3: 8.5
```