Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

Third Year B. Tech (EL & CE)

Semester: VI **Subject:** Data Science for Engineering

Name: Shreerang Mhatre **Class: TY**

Roll No: 52 **Batch: A2**

Experiment No: 04

Name of the Experiment: **Understanding Statistical concepts in Python.**

Performed on: 26/02/2024

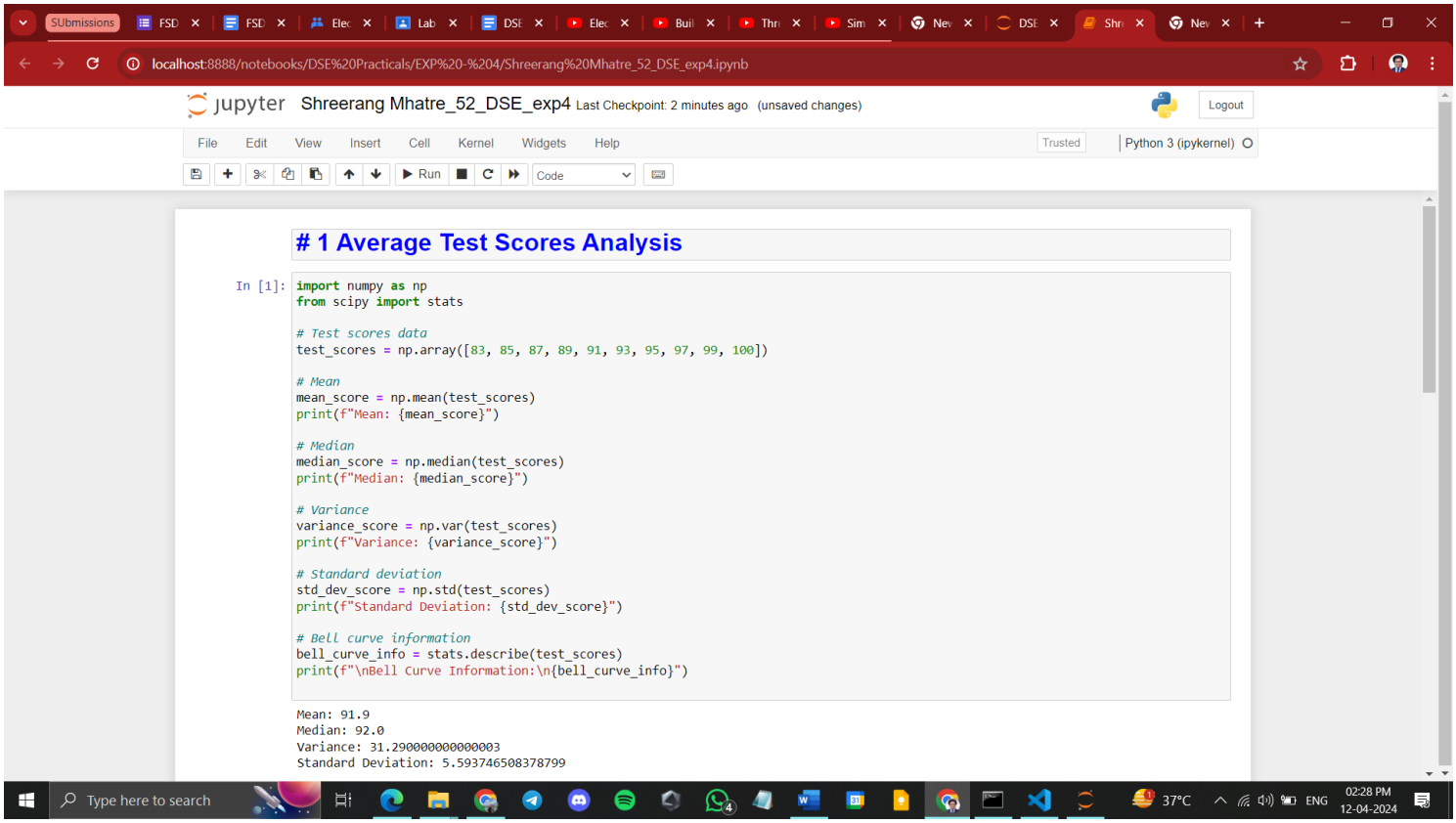
Submitted on: 18/04/2024

Problem Statement:

The average test scores are given: test scores: 83,85,87,89,91,93,95,97,99,100. Find Mean, Median, Variance, Standard deviation of the data. Show the information on the bell curve.
Consider given product price data: price_data= [13,43,54,34,40,56,34,61,34,23]. Find Range, 25th Percentile and IQR.
A person tries to analyse the last 12 months interest rate of the investment firm to understand the risk factor for the future investment. The interest rates are:12.05, 13, 11, 18, 10, 11.5, 15.08, 21, 6, 8, 13.2, 7.5.

Calculate Skewness and Kurtosis and comment on it.

1. Hypothesis Testing
a. Consider below data and tests whether a data sample has a Gaussian distribution by formulating hypothesis test
data = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]



jupyter Shreerang Mhatre_52_DSE_exp4 Last Checkpoint: 2 minutes ago (unsaved changes)

Logout

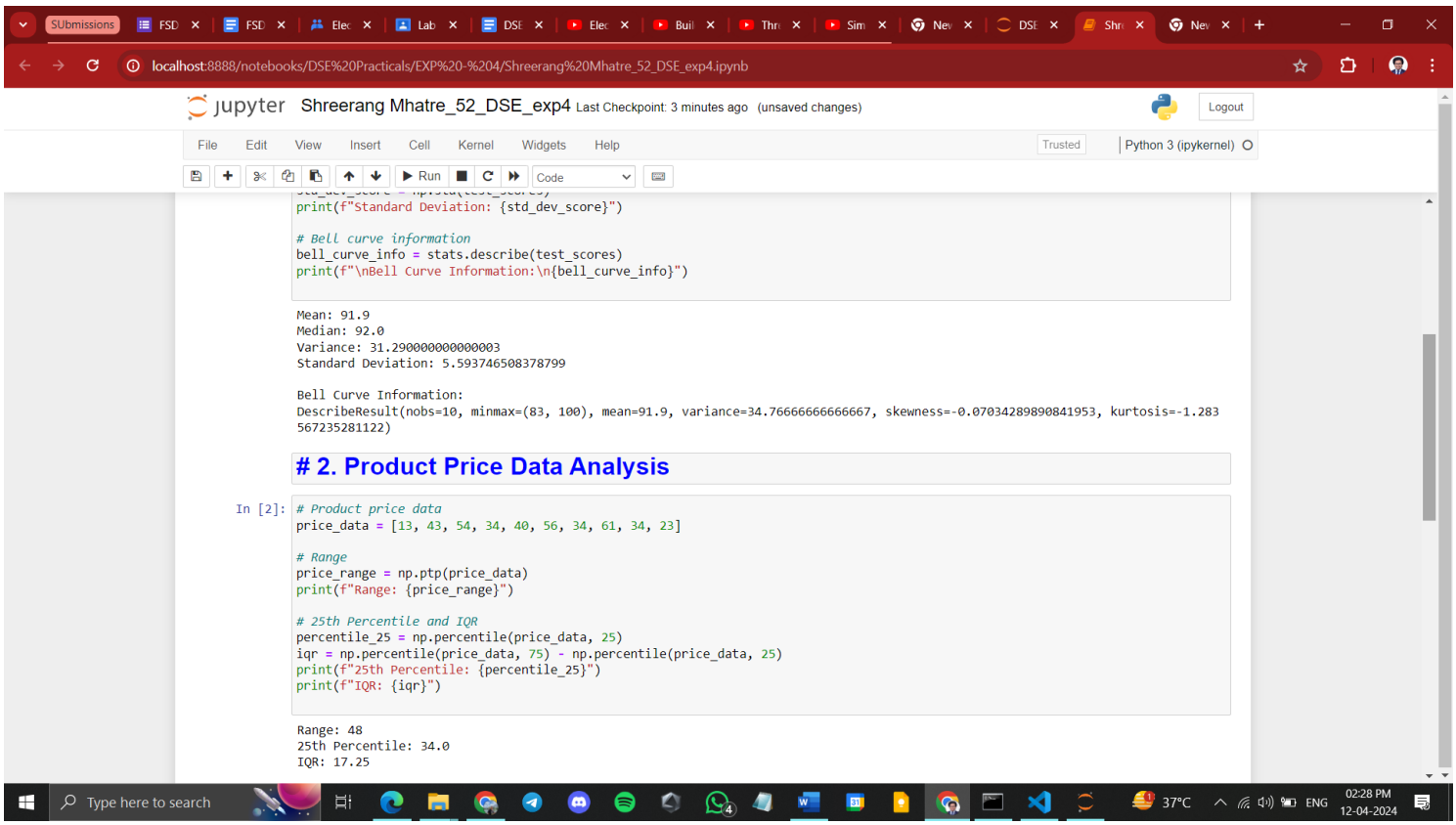
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

1 Average Test Scores Analysis

In [1]:
```
import numpy as np
from scipy import stats

# Test scores data
test_scores = np.array([83, 85, 87, 89, 91, 93, 95, 97, 99, 100])

# Mean
mean_score = np.mean(test_scores)
print(f"Mean: {mean_score}")

# Median
median_score = np.median(test_scores)
print(f"Median: {median_score}")

# Variance
variance_score = np.var(test_scores)
print(f"Variance: {variance_score}")

# Standard deviation
std_dev_score = np.std(test_scores)
print(f"Standard Deviation: {std_dev_score}")

# Bell curve information
bell_curve_info = stats.describe(test_scores)
print(f"\nBell Curve Information:\n{bell_curve_info}")
```

```
Mean: 91.9
Median: 92.0
Variance: 31.290000000000003
Standard Deviation: 5.593746508378799
```



jupyter Shreerang Mhatre_52_DSE_exp4 Last Checkpoint: 3 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

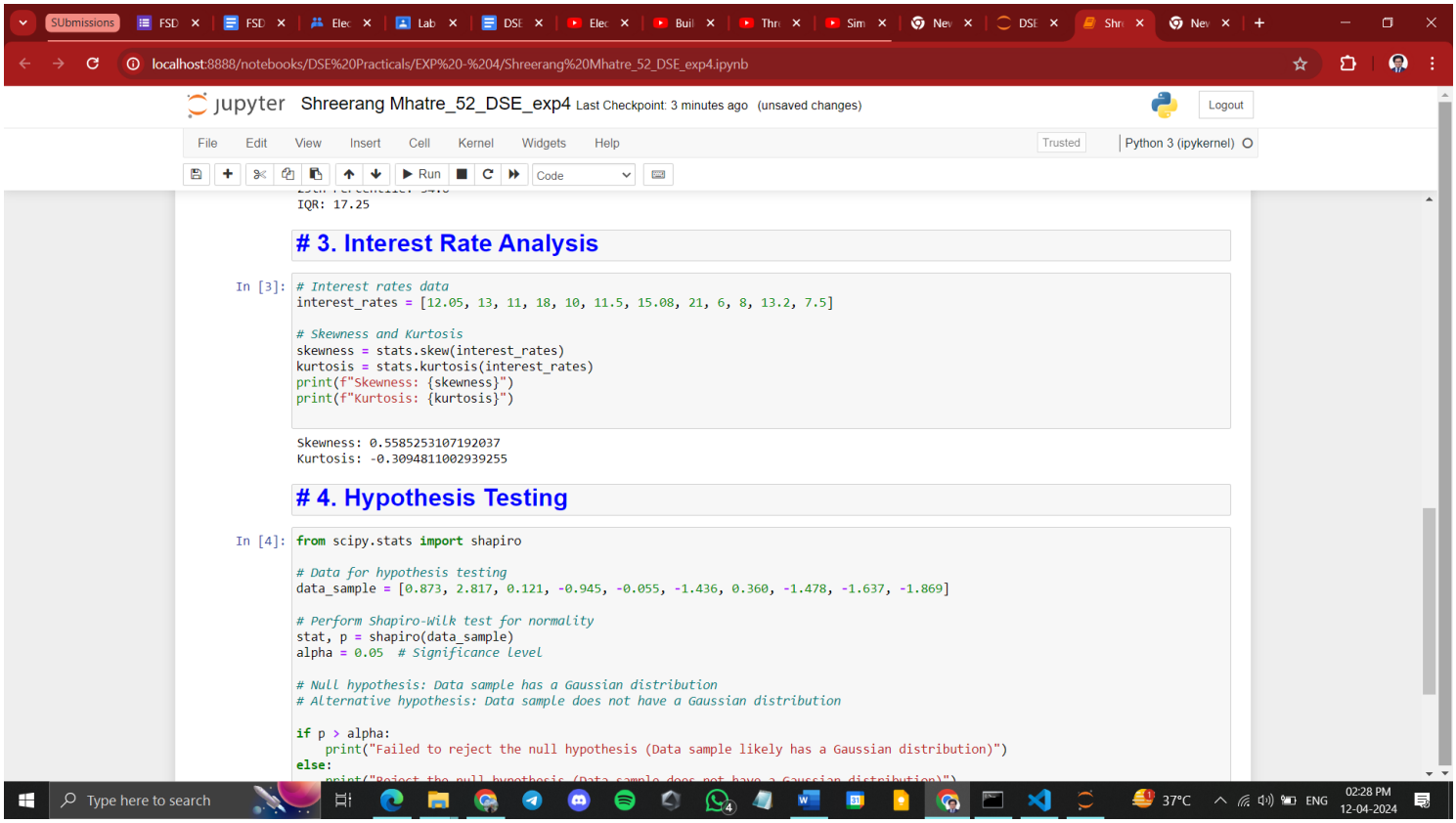
```
print(f"Standard Deviation: {std_dev_score}")

# Bell curve information
bell_curve_info = stats.describe(test_scores)
print(f"\nBell Curve Information:\n{bell_curve_info}")
```

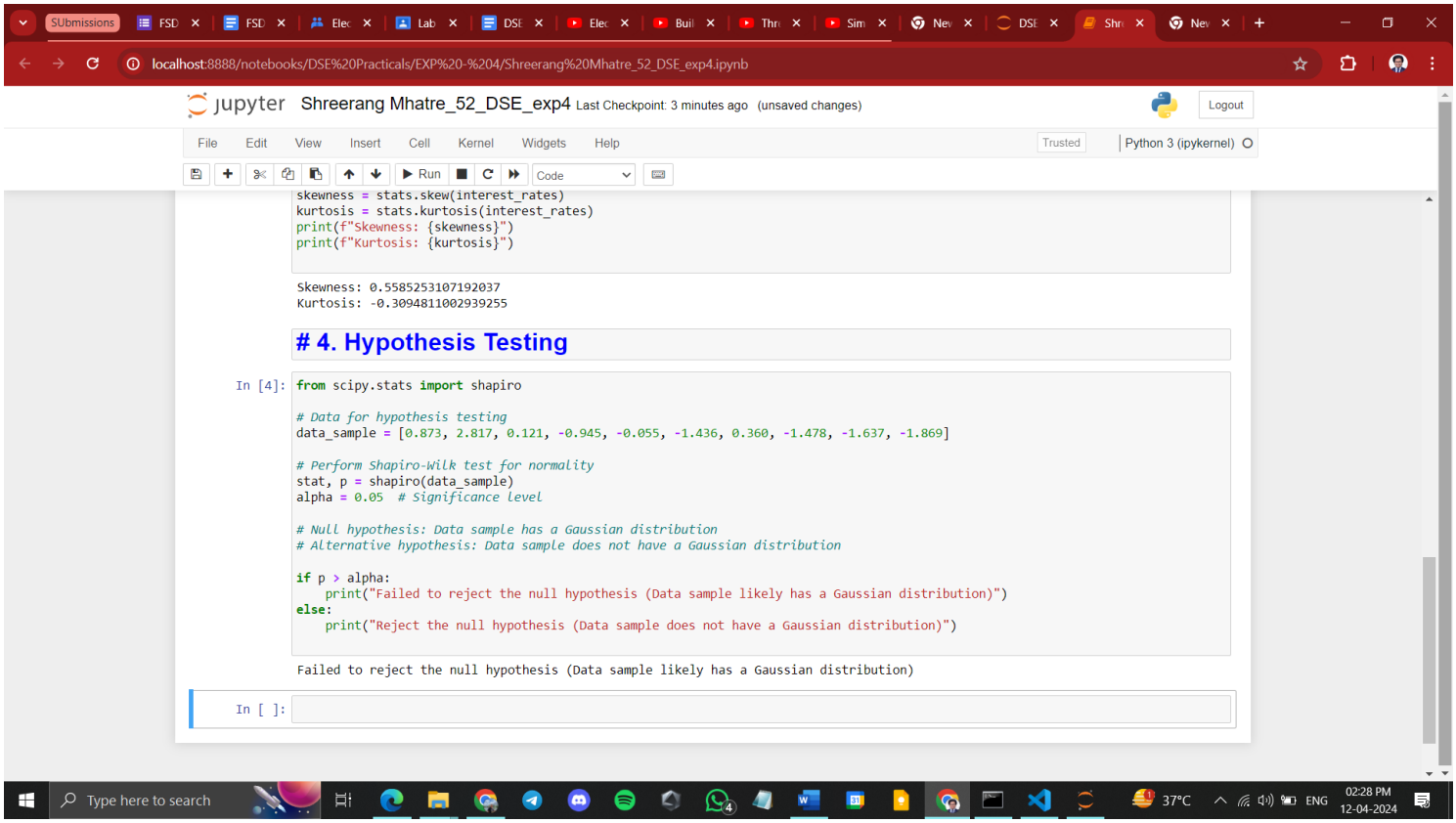
```
Mean: 91.9
Median: 92.0
Variance: 31.290000000000003
Standard Deviation: 5.593746508378799

Bell Curve Information:
DescribeResult(nobs=10, minmax=(83, 100), mean=91.9, variance=34.76666666666667, skewness=-0.07034289890841953, kurtosis=-1.283
567235281122)
```

2. Product Price Data Analysis

In [2]:
```
# Product price data
price_data = [13, 43, 54, 34, 40, 56, 34, 61, 34, 23]

# Range
price_range = np.ptp(price_data)
print(f"Range: {price_range}")

# 25th Percentile and IQR
percentile_25 = np.percentile(price_data, 25)
iqr = np.percentile(price_data, 75) - np.percentile(price_data, 25)
print(f"25th Percentile: {percentile_25}")
print(f"IQR: {iqr}")
```

```
Range: 48
25th Percentile: 34.0
IQR: 17.25
```


localhost:8888/notebooks/DSE%20Practicals/EXP%20-%204/Shreerang%20Mhatre_52_DSE_exp4.ipynb
jupyter Shreerang Mhatre_52_DSE_exp4 Last Checkpoint: 3 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
IQR: 17.25
3. Interest Rate Analysis
In [3]:
# Interest rates data
interest_rates = [12.05, 13, 11, 18, 10, 11.5, 15.08, 21, 6, 8, 13.2, 7.5]
# Skewness and Kurtosis
skewness = stats.skew(interest_rates)
kurtosis = stats.kurtosis(interest_rates)
print(f"Skewness: {skewness}")
print(f"Kurtosis: {kurtosis}")
Skewness: 0.5585253107192037
Kurtosis: -0.3094811002939255
4. Hypothesis Testing
In [4]:
from scipy.stats import shapiro
# Data for hypothesis testing
data_sample = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
# Perform Shapiro-Wilk test for normality
stat, p = shapiro(data_sample)
alpha = 0.05 # Significance level
# Null hypothesis: Data sample has a Gaussian distribution
# Alternative hypothesis: Data sample does not have a Gaussian distribution
if p > alpha:
print("Failed to reject the null hypothesis (Data sample likely has a Gaussian distribution)")
else:
Type here to search
37°C
ENG
02:28 PM
12-04-2024


localhost:8888/notebooks/DSE%20Practicals/EXP%20-%204/Shreerang%20Mhatre_52_DSE_exp4.ipynb
jupyter Shreerang Mhatre_52_DSE_exp4 Last Checkpoint: 3 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
skewness = stats.skew(interest_rates)
kurtosis = stats.kurtosis(interest_rates)
print(f"Skewness: {skewness}")
print(f"Kurtosis: {kurtosis}")
Skewness: 0.5585253107192037
Kurtosis: -0.3094811002939255
4. Hypothesis Testing
In [4]:
from scipy.stats import shapiro
# Data for hypothesis testing
data_sample = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
# Perform Shapiro-Wilk test for normality
stat, p = shapiro(data_sample)
alpha = 0.05 # Significance level
# Null hypothesis: Data sample has a Gaussian distribution
# Alternative hypothesis: Data sample does not have a Gaussian distribution
if p > alpha:
print("Failed to reject the null hypothesis (Data sample likely has a Gaussian distribution)")
else:
print("Reject the null hypothesis (Data sample does not have a Gaussian distribution)")
Failed to reject the null hypothesis (Data sample likely has a Gaussian distribution)
In [ ]:
Type here to search
37°C
ENG
02:28 PM
12-04-2024

Exp 4

Shreerang Mhatre 52 A2

* Post Lab Questions :

Q1) How do measures of central tendency (mean, median, mode) differ from eachother?

→ (1) Mean -

The mean, also known as the average, is calculated by adding up all the values in a dataset and then dividing by the number of values. It represents the central value around which the data points are distributed.

(2) Median -

The median is the middle value in a sorted list of data. If there is an odd number of data points, the median is the middle value.

(3) Mode -

The mode is the value that appears most frequently in a dataset. A dataset can have multiple modes if more than one value occurs with the highest frequency.

Q2) Discuss the relationship between variance and standard deviation. How are they related mathematically?

→ ① Variance -

Variance measures the spread or dispersion of a set of data points around the mean. It is calculated by taking the average of the squared differences between each data point & mean.

② Standard Deviation -

Standard deviation is the square root of the variance. It measures the average distance of data points from the mean. standard deviation is often preferred over variance because it is in the same units as the original data.

$$\text{standard deviation}(\sigma) = \sqrt{\text{variance}(\sigma^2)}$$

Q3) What are skewness and kurtosis? How do their values affect the shape of a distribution?

→ ① Skewness -

Skewness measures the asymmetry of the distribution of data points around the mean. A distribution can be positively skewed, negatively skewed or approximately symmetric. Skewness quantifies the lack of symmetry in the data distribution.

② Kurtosis -

Kurtosis measures the 'tailedness' of the distribution, indicating how much data is in the tails compared to the center of the distribution. A distribution with high kurtosis has more data in the tails, while a distribution with low kurtosis has less data in the tails.