Question 1

You perform a clinical trial to study new drug. You have 20 volunteers with some disease. You randomly split all the volunteers into two groups (10 volunteers in each): the treatment group and the control group. Volunteers in the treatment group receive the new drug, volunteers in the control group receive placebo (pills that looks like a drug but do not have active substance). You conclude that new drug is effective if people who take the drug will recover faster (on average) than people in the control group. If your drug is effective, you will invest in its production, otherwise you will look for another drug. Assume that you obtained the following data (disease duration in days).

control group	treatment group
6	7
7	6
7	6
5	5
7	5
8	6
8	7
7	5
7	5
7	8

Describe this problem in terms of statistical hypothesis testing framework. How would you model your data in terms of random variables? State the null hypothesis and the alternative. Will your alternative be one-sided or two-sided? Why? What kind of statistical test will you use? Why this test? Use this test (apply Python if necessary and provide your code), analyse the results and provide a conclusion in mathematical and real-life terms. Would you invest into production of this drug?

In terms of statistical hypothesis testing framework, the problem is to determine if there is a statistically significant difference in recovery time between the treatment group and the control group. Specifically, in this case we are interested faster recovery time for the treatment group.

Considering the problem, we can model the data using random variables as follows:
Let X be a random variable representing the recovery time for a volunteer in the treatment group and Y be a random variable representing the recovery time for a volunteer in the control group.
The recovery times for individuals in both groups can be considered as independent and identically distributed (i.i.d) random variables, assuming that each individual's recovery time is not influenced by others in the group.

Having collected the observed recovery time data for each group, we obtain two resulting sets of observations:
$X = \{x_1, x_2, ..., x_{10}\}$ representing the recovery times for the treatment group.
$Y = \{y_1, y_2, ..., y_{10}\}$ representing the recovery times for the control group.

The goal is to analyze the observed data sets and determine if there is a statistically significant difference in recovery time between the treatment group and the control group, using appropriate statistical tests.

Null Hypothesis (H_0): There is no difference in recovery time between the treatment group and the control group.
Alternative Hypothesis (H_1): The new drug is effective, and the average recovery time for the treatment group is shorter than the average recovery time for the control group.
Mathematically, it can be expressed as:
$H_0: \mu X = \mu Y$
$H_1: \mu X < \mu Y$
where μX and μY are the values of population mean recovery time for the treatment group and for the control group correspondingly.

Based on the condition that the new drug is considered effective only if people taking the drug recover faster than people in the control group, the alternative hypothesis can be formulated as a one-sided alternative. In this scenario we are specifically interested in faster recovery time, as it determines the decision to invest in the production.

The appropriate test to apply for this purpose is the 2-sample t-test. It is used to compare the means of two independent groups, which is precisely what we want to do in this case, as we have two groups of volunteers, and we want to compare their mean recovery times to determine if there is a statistical significance of the observed difference.

Python code for the 2-sample t-test implementation:

```
import numpy as np
from scipy.stats import ttest_ind

treatment_group = np.array([7, 6, 6, 5, 5, 6, 7, 5, 5, 8])
control_group = np.array([6, 7, 7, 5, 7, 8, 8, 7, 7, 7])

t_statistic, p_value = ttest_ind(treatment_group, control_group, alternative='less')

print(f't-statistic: {t_statistic}\np-value: {p_value}')
```

The output:

```
t-statistic: -2.076923076923078
p-value: 0.026201003203565984
```

Interpreting the results:

The t-statistic is -2.077. It indicates the magnitude of the difference between the treatment group and the control group mean recovery times. The negative value suggests that the treatment group has a lower average recovery time compared to the control group.
The p-value is 0.026, which is less than the common significance level of 0.05. Therefore, we can reject the null hypothesis at the 0.05 level of significance.

In practical terms, based on the analysis, there is sufficient evidence to conclude that the new drug has a statistically significant effect on reducing recovery time compared to the placebo, as, on average, the volunteers who received the new drug experienced a shorter recovery time compared to those who received the placebo.

This finding gives a positive support to a decision regarding investment in the production of the drug, as it shows promising results in reducing recovery time compared to the placebo.