Problem1 Statement:

Cold Storage started its operations in Jan 2016. They are in the business of storing Pasteurized Fresh Whole or Skimmed Milk, Sweet Cream, Flavoured Milk Drinks. To ensure that there is no change of texture, body appearance, separation of fats the optimal temperature to be maintained is between 2 and 4 C. In the first year of business they outsourced the plant maintenance work to a professional company with stiff penalty clauses. It was agreed that if the it was statistically proven that probability of temperature going outside the 2 - 4 C during the one-year contract was above 2.5% and less than 5% then the penalty would be 10% of AMC (annual maintenance case). In case it exceeded 5% then the penalty would be 25% of the AMC fee. The average temperature data at date level is given in the file “Cold_Storage_Temp_Data.csv”

Exploratory Data Analysis:

Let us now look at the first 5 rows of the data to understand how data looks.

	Season	Month	Date	Temperature
0	Winter	Jan	1	2.3
1	Winter	Jan	2	2.2
2	Winter	Jan	3	2.4
3	Winter	Jan	4	2.8
4	Winter	Jan	5	2.5

Dataset has 4 variables Season, Month, Date and Temperature. Season and Month both are categorical columns while Date is integer and Temperature is a float type.

Descriptive Statistics for the dataset:

	Season	Month	Date	Temperature
count	365	365	365.000000	365.000000
unique	3	12	NaN	NaN
top	Winter	May	NaN	NaN
freq	123	31	NaN	NaN
mean	NaN	NaN	15.720548	3.002466
std	NaN	NaN	8.808321	0.465832
min	NaN	NaN	1.000000	1.700000

25%	NaN	NaN	8.000000	2.700000
50%	NaN	NaN	16.000000	3.000000
75%	NaN	NaN	23.000000	3.300000
max	NaN	NaN	31.000000	4.500000

There are 3 unique values in Season column in which winter season has the most values in the season column. Mean for the Temperature is 3.002 with the standard deviation of 0.466.
Check for Null values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 365 entries, 0 to 364
Data columns (total 4 columns):
Season          365 non-null object
Month           365 non-null object
Date            365 non-null int64
Temperature     365 non-null float64
dtypes: float64(1), int64(1), object(2)
memory usage: 11.5+ KB
```

From the above results, it is evident that there is no null values present in the dataset.

1. **Find mean cold storage temperature for Summer, Winter and Rainy Season?**

   Summer Mean = sum (Temperature of summer)/number of rows for summer Winter Mean = sum (Temperature of Winter)/number of rows for winter Rainy Mean = sum (Temperature of Rainy)/number of rows for Rainy

   The mean cold storage temperature for Rainy is **3.087705** The mean cold storage temperature for summer is **3.147500** The mean cold storage temperature for winter is **2.776423**

2. **Find overall mean for the full year?**

   Overall Mean = sum (Temperature)/number of rows

   Overall mean for the full year is **3.002**

3. **Find Standard Deviation for the full year?**

   Overall Standard Deviation for the full year is **0.466**

4. **Assume Normal distribution, what is the probability of temperature having fallen below 2 C?**

μ = overall mean for the full year = overall_mean = **3.002**
σ = standard deviation for the full year = overall_std = **0.466**

x1 = lower bound = 2 degree Celsius

Using the z-score formula = (x1-μ)/σ = (2-3.002)/0.466 = -2.151

Now using the z-table, the probability of temperature having fallen below 2 C is **0.0156**

5. **Assume Normal distribution, what is the probability of temperature having gone above 4 C?**

   μ = overall mean for the full year = overall_mean = **3.002**
   σ = standard deviation for the full year = overall_std = **0.466**

   x2= upper bound = 4 degree Celsius

   Using the z-score formula = (x2-μ)/σ = (4-3.002)/0.466 = 2.141

   Now using the z-table, the probability of temperature having gone above 4 C is **0.0161**

6. **What will be the penalty for the AMC Company?**

   Penalty for the AMC company is applicable when the probability of temperature is going out of 2^o – 4^o Celsius.

   Total probability of temperature going out of the above range = probability of temperature having gone above 4^o C + the probability of temperature having below 2^o C

   Total probability = 0.0318

   Since the probability of temperature going outside the 2^o C – 4^o C during the one-year contract was above 2.5% and less than 5% then the penalty would be **10%**

Problem 1: Summary

1. Mean cold storage temperature for
   a. Summer: 3.147
   b. Winter: 2.7764
   c. Rainy: 3.08770
2. Overall mean for the full year: 3.002
3. Standard Deviation for the full year: 0.466
4. Probability of temperature having fallen below 2^o C: 0.0157
5. Probability of temperature having gone above 4^o C: 0.016
6. Penalty for the AMC Company: 10%

Problem2 Statement:

In Mar 2018, Cold Storage started getting complaints from their Clients that they have been getting complaints from end consumers of the dairy products going sour and often smelling. On getting these complaints, the supervisor pulls out data of last 35 days' temperatures. As a safety measure, the Supervisor decides to be vigilant to maintain the temperature 3.9 C or below.

Assume 3.9 C as upper acceptable value for mean temperature and at alpha = 0.1 do you feel that there is need for some corrective action in the Cold Storage Plant or is it that the problem is from procurement side from where Cold Storage is getting the Dairy Products. The data of the last 35 days is in "Cold_Storage_Mar2018.csv"

Exploratory Data Analysis:

	Season	Month	Date	Temperature
0	Summer	Feb	11	4.0
1	Summer	Feb	12	3.9
2	Summer	Feb	13	3.9
3	Summer	Feb	14	4.0
4	Summer	Feb	15	3.8

Descriptive Statistics for the dataset:

	Season	Month	Date	Temperature
count	35	35	35.000000	35.000000
unique	1	2	NaN	NaN
top	Summer	Feb	NaN	NaN
freq	35	18	NaN	NaN
mean	NaN	NaN	14.400000	3.974286
std	NaN	NaN	7.389181	0.159674
min	NaN	NaN	1.000000	3.800000
25%	NaN	NaN	9.500000	3.900000
50%	NaN	NaN	14.000000	3.900000
75%	NaN	NaN	19.500000	4.100000
max	NaN	NaN	28.000000	4.600000

There are 3 unique values in Season column in which summer season has the most values in the season column. Mean for the Temperature is 3.97 with the standard deviation of 0.16.

Check for Null values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35 entries, 0 to 34
Data columns (total 4 columns):
Season           35 non-null object
Month            35 non-null object
Date             35 non-null int64
Temperature     35 non-null float64
dtypes: float64(1), int64(1), object(2)
Memory usage: 1.2+ KB
```

From the above results, it is evident that there is no null values present in the dataset.

1. **Which Hypothesis test shall be performed to check if corrective action is needed at the cold storage plant? Justify your answer.**

There are 2 tests, z-test and t-test which are generally used to perform for hypothesis testing.

A z-test is a statistical test used to determine whether two population means are different when the population variances are known and the sample size is large. The test statistic is assumed to have a normal distribution.

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the dataset follows a normal distribution and have unknown population variances. A t- test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.

Since we do not have the information about the population standard deviation, we will perform t-test in this case

2. **State the Hypothesis, perform hypothesis test and determine p-value?**

A **null hypothesis** is a hypothesis that says there is no statistical significance between the two variables in the hypothesis. It is the hypothesis that the researcher is trying to disprove.

An **alternative hypothesis** simply is the inverse, or opposite, of the null hypothesis.

As a safety measure, the task is to maintain the temperature 3.9 C or below. So the hypothesis would as per below:

Null Hypothesis: *Temperature is maintained equal to 3.9°C and hence no corrective action is required.*

H_0: $\mu = 3.9^{o}$ C

Alternate Hypothesis: *Temperature is not equal to 3.9^o C and hence corrective action in the Cold Storage Plant is required.*

H_A: μ != 3.9^o C

First decide the level of significance:

The level of significance is defined as the probability of rejecting a null hypothesis by the test when it is really true, which is denoted as α. That is, P (Type I error) = α. Confidence level: The level of significance 0.1 is related to the 90% confidence level.

Level of Significance α = 0.1

From the t- test performed, we got the below results:

One sample t test
t statistic: 2.752358609800241 p value: 0.009422395404264433

At the level of 10% significance, p-value = 0.009. Since p-value < 0.1, we will reject the n ull hypothesis.
We have evidence to reject the null hypothesis

3. Give your inference?

Using t Test: Conclusion

- t stats = 2.75 and p-value = 0.0094
- The P-value is less than 0.1, indicating that it is highly likely that these results would be observed under the alternate hypothesis.
- We fail to accept H_o.
- Thus, we accept H_a given this result.
- Temperature is above *3.9^o C* and hence corrective action in the Cold Storage Plant is required.

Final Conclusion:

Because we did not have the population standard deviation given, we have to go with T-test in this case, which states that probability of the cold storage being at fault is statistically more responsible for sour and often smelling dairy products