Problem 1 Statement

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The data on hours of relief can be found in the following .csv file: Fever.csv

[Assume all of the ANOVA assumptions are satisfied]

Exploratory Data Analysis :

A	B	Volunteer	Relief
1	1	1	2.4
1	1	2	2.7
1	2	2	4.2
2	1	1	5.8
2	1	2	5.2
2	1	3	5.5
3	2	3	10.6
3	2	4	10.1
3	3	1	13.5
3	3	2	13
3	3	3	13.3
3	3	4	13.2

We are provided with the above data set of 36 rows and 4 columns. Even though columns 'A' ,'B' and 'Volunteer' have integer type data . We have converted all of these columns into categorical type.

Therefore ,of the available columns 3 are categorical type and 1 is float type. The data has no Null values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   A          36 non-null     category
 1   B          36 non-null     category
 2   Volunteer  36 non-null     category
 3   Relief     36 non-null     float64
dtypes: category(3), float64(1)
memory usage: 924.0 bytes
```

Descriptive statistics for the dataset:

	A	B	Volunteer	Relief
count	36.0	36.0	36.0	36.000000
unique	3.0	3.0	4.0	NaN
top	3.0	3.0	4.0	NaN
freq	12.0	12.0	9.0	NaN
mean	NaN	NaN	NaN	7.183333
std	NaN	NaN	NaN	3.272090
min	NaN	NaN	NaN	2.300000
25%	NaN	NaN	NaN	4.675000
50%	NaN	NaN	NaN	6.000000
75%	NaN	NaN	NaN	9.325000
max	NaN	NaN	NaN	13.500000

We have Columns **'A' , 'B'** and **'Volunteer'** as **categorical type** data and column **'Relief'** as **Float type.**

There are **3** unique values in columns **'A' ,'B'** with a frequency value of **12** and there are **4** unique values in column **'Volunteer'** with a frequency value of **9** .

As per the details resulted from the descriptive statistics of the dataset, we can find that:

All the four variables have a **value count** of **36** with **no null values**. The column **'Relief'** has **float type** data with following values of **mean** = 7.183333 , **median** =6.000 , and **max** = 13.500000

1.1. State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'A' and 'B' individually. [both statement and statistical form like Ho=mu, Ha>mu]

Variable – **'A'**
Null Hypothesis (Ho) : **mu1 = mu2= mu3**
The means of 'Relief' variable with respect to each level of ingredient variable 'A' are same
Alternate Hypothesis (Ha) : **mu1 =/ mu2 =mu3** or **mu1 = mu2 /= mu3** or **mu1=mu3/=mu2** or **mu1 /= mu2 /= mu3** At least one of the mean of 'Relief' variable with respect to each level of variable 'A' is not same

Variable – **'B'**
Null Hypothesis (Ho) : **mu1 = mu2= mu3**
The means of 'Relief' variable with respect to each level of ingredient variable 'B' are same
Alternate Hypothesis (Ha) : **mu1 =/ mu2 =mu3** or **mu1 = mu2 /= mu3** or **mu1=mu3/=mu2** or **mu1 /= mu2 /= mu3** At least one of the mean of 'Relief' variable with respect to each level of variable 'B' is not same

1.2. Perform one-way ANOVA for variable 'A' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.02	110.010000	23.465387	4.578242e-07
Residual	33.0	154.71	4.688182	NaN	NaN

As per above result , we can see that the value of p is less the level of significance (0.05) .

Therefore we reject the null hypothesis, that means at least one of the mean of 'Relief' variable with respect to each level of variable 'A' is not same.

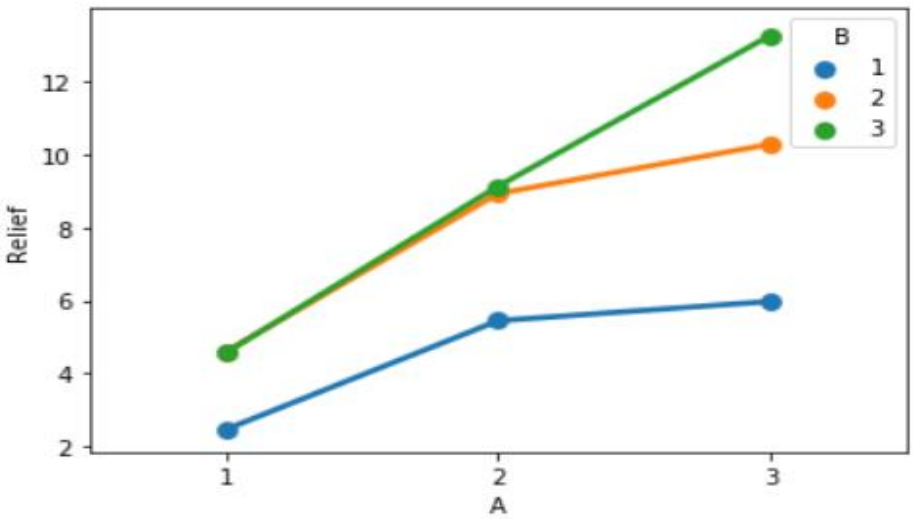
1.3. Perform one-way ANOVA for variable 'B' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

	df	sum_sq	mean_sq	F	PR(>F)
C(B)	2.0	123.66	61.830000	8.126777	0.00135
Residual	33.0	251.07	7.608182	NaN	NaN

As per above result , we can see that the value of p is less than the level of significance (0.05) .

Therefore we reject the null hypothesis, that means at least one of the mean of 'Relief' variable with respect to each level of variable 'B' is not same.

1.4. Analyse the effects of one variable on another with the help of an interaction plot. What is the interaction between the two treatments? [hint: use the 'pointplot' function from the 'seaborn' function]



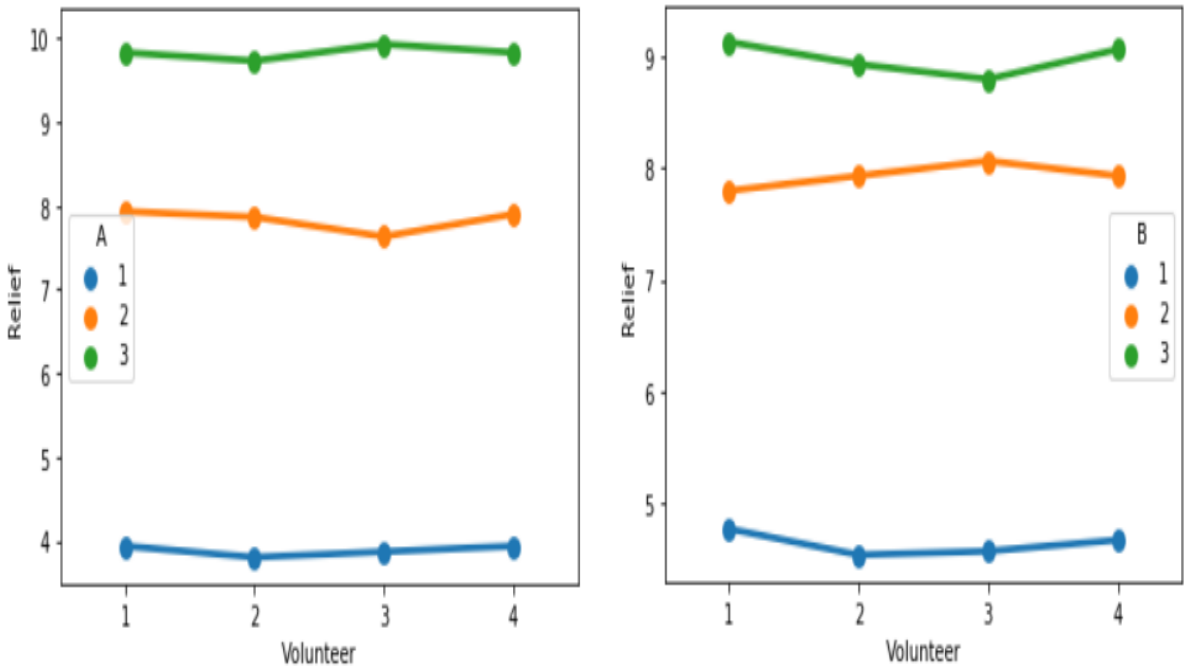
As per plot, We can infer that Relief increase with the increase in level of Variable 'A' from level 1 to 3. Also for a given a level of variable 'A' , there is a higher Relief with increase in level of Variable 'B' from level 1 to 3.

From above plot , we can find that there is interaction between the two categorical variables 'A' & 'B'

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.020	110.010000	1827.858462	1.514043e-29
C(B)	2.0	123.660	61.830000	1027.329231	3.348751e-26
C(A):C(B)	4.0	29.425	7.356250	122.226923	6.972083e-17
Residual	27.0	1.625	0.060185	NaN	NaN

We can see from above result that p value for interaction between variable 'A' and variable 'B' is less than the level of significance (0.05).

Therefore , we reject null hypothesis, that means there is significant interaction between the two treatments.



From the above plot, we can see that for a given Volunteer, there is increase in Relief with increase in level of variables in both 'A' & 'B'.

Relief value is almost at same level for different volunteer in both variables 'A' & 'B'.

C(A):C(Volunteer)	6.0	0.191111	0.031852	0.370889	8.837392e-01
C(B):C(Volunteer)	6.0	0.331111	0.055185	0.642588	6.954884e-01
Residual	12.0	1.030556	0.085880	NaN	NaN

Since the p value is greater than 0.05 , therefore we can say that there is almost no interaction between the variables 'A' and 'Volunteer' and also similarly there is almost no interaction between the variables 'A' and 'Volunteer'.

1.5. Perform a two-way ANOVA based on the different ingredients (variable 'A' & 'B' along with their interaction 'A*B') with the variable 'Relief' and state your results.

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.020000	110.010000	1280.978976	1.026783e-14
C(B)	2.0	123.660000	61.830000	719.961186	3.187330e-13
C(Volunteer)	3.0	0.072222	0.024074	0.280323	8.385898e-01
C(A):C(B)	4.0	29.425000	7.356250	85.657682	1.020340e-08
C(A):C(Volunteer)	6.0	0.191111	0.031852	0.370889	8.837392e-01
C(B):C(Volunteer)	6.0	0.331111	0.055185	0.642588	6.954884e-01
Residual	12.0	1.030556	0.085880	NaN	NaN

From the above table, we can infer the following results,

Variable 'A' has p value less than 0.05 , So it plays significant role in relation with variable 'Relief'.

Variable 'B' has p value less than 0.05 , So it plays significant role in relation with variable 'Relief'.

Variables 'C' has p value greater than 0.05 , So it plays almost no interaction with variable 'Relief.

P value of interaction between variables 'A' and 'B' is less than 0.05, that means there is significant interaction between the two variables.

1.6. Mention the business implications of performing ANOVA for this particular case study

We use ANOVA test to determine whether different populations are statistically different from each other. It gives us single number(the f- statistic) and one p-value to help us support or reject the null hypothesis.

Based on the results from the ANOVA test , We can see that both the ingredients 'A' &'B' in the compound for the relief play a significant role in deciding the hours required for relief to be felt by the patient. Also there is very high interaction between the two ingredients with respect to variable 'Relief'.

Moreover , we can see that combination of level 1 of both ingredients 'A' & 'B' resulted in lower hours required for relief and the combination of level 3 of both ingredients 'A' & 'B' resulted in higher hours required for relief.