Problem 1 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

Step 1 - Importing all the necessary library

Step 2 - Loading the file and finding the information of the data *

Before doing ANOVA test finding whether mean can be used to intepret the data. Hence mean of 3 active levels are calcualted by groupby function

The mean relief of 1 active level ranges from 3.80 to 3.93

The mean relief of 2 active level ranges from 7.63 to 7.93

The mean relief of 3 active level ranges from 9.73 to 9.93

By using describe function mean, count, min, max, ,median, 25% and 75% is calculated

A B Volume Relief count 36.000000 36.000000 36.000000 36.000000 mean 2.000000 2.000000 2.500000 7.183333 std 0.828079 0.828079 1.133893 3.272090 min 1.000000 1.000000 1.000000 2.300000 25% 1.000000 1.000000 1.750000 4.675000 50% 2.000000 2.000000 2.500000 6.000000 75% 3.000000 3.000000 3.250000 9.325000 max 3.000000 3.000000 4.000000 13.500000

Step 3 - By using info , check if there is any null value and there are no null values

Step 4 - Categorical function is used to find the distribution of data in each category

Average alone is not good enough description of the data, though there is quite some variation in relief. For example, the original data show that the minimum relief time of 3 active level is in the range of maximum relief time of 2 active level in A

The ANOVA model helps in estimating the total amount of variation that exists in the relief time. Hence split the total variation into two: Between and Within Group variation for A

ANOVA: Hypothesis test Is one active component has higher relief time, or it is random noise due to sampling effect

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]

1. Null hypothesis (Ho): There is no difference in average relief for any of the active level in A

   ```
   HoA: Mu1 = Mu2 = Mu3
   ```

   Alternative hypothesis (Ha): There is a difference in average relief for any of the active level in A

   ```
   HaA: Mu1 not equal to Mu2 not equal to Mu3
   ```

2. Null hypothesis (Ho): There is no difference in average relief for any of the active level in B

   ```
   HoB: Mu1 = Mu2 = Mu3
   ```

   Alternative hypothesis (Ha): There is a difference in average relief for any of the active level in B

   ```
   HaB: Mu1 not equal to Mu2 not equal to Mu3
   ```

3. Null hypothesis (Ho): The effect of one independent variable A on average relief does not depend on the effect of the other independent variable B. Alternative hypothesis (Ha): There is an interaction effect between A and B

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.

One way ANOVA F test using Statsmodels

To perform ANOVA below formula is used

formula = 'Relief ~ C(A) model = ols(formula, DF).fit() aov_table = anova_lm(model) print(aov_table)

Step 1

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

-Df shows the degrees of freedom for each variable (number of levels in the variable minus 1). -Sum sq is the sum of squares (a.k.a. the variation between the group means created by the levels of the independent variable and the overall mean). -Mean sq shows the mean sum of squares (the sum of squares divided by the degrees of freedom). -F value is the test statistic from the F-test (the mean square of the variable divided by the mean square of each parameter). -Pr(>F) is the p-value of the F statistic, and shows how likely it is that the F-value calculated from the F-test would have occurred if the null hypothesis of no difference was true.
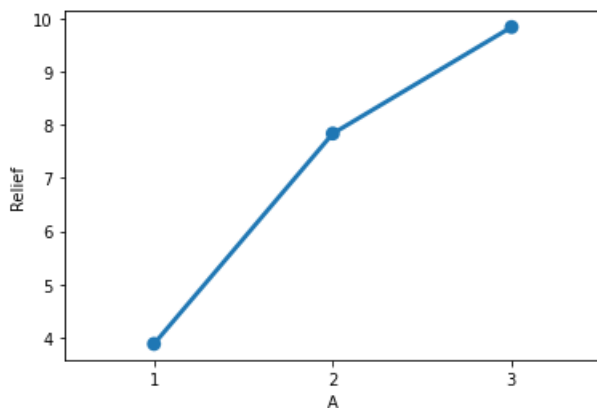
From this output we can see that 1.Amount of active ingredients explain a significant amount of variation in average relief (p-values < 0.05). Hence reject the null hypothesis. This means that with the observed data, there is enough evidence to assume a general difference in the relief time of the active ingredients(1,2,3) in A

Step 2 - Plotting the graph

In [29]:

Out[29]:

```
<AxesSubplot:xlabel='A', ylabel='Relief'>
```


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.

**Step 1 - One way ANOVA F test using Statsmodels

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

From this output we can see that 1.Amount of active ingredients explain a significant amount of variation in average relief (p-values < 0.05). Hence reject the null hypothesis. This means that with the observed data, there is enough evidence to assume a general difference in the relief time of the active ingredients(1,2,3) in B
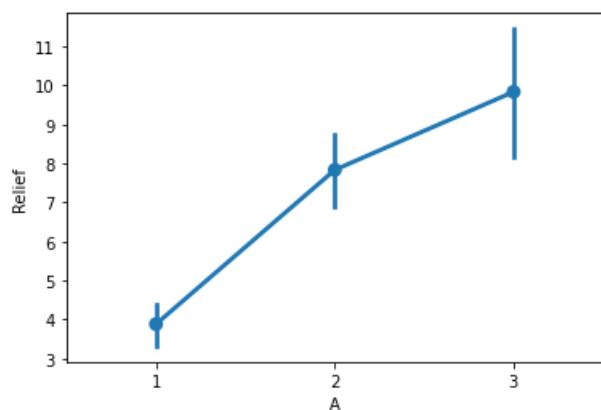
In [11]:

Step 2 - Plotting the graph
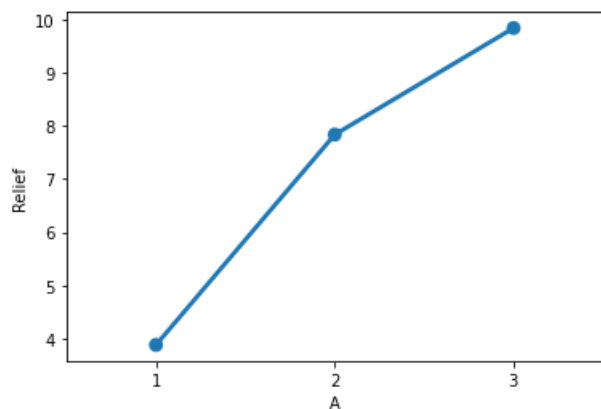
```
<AxesSubplot:xlabel='A', ylabel='Relief'>
```

```
<AxesSubplot:xlabel='A', ylabel='Relief'>
```


1.4) Analyse the effects of one variable on another with the help of an interaction plot.
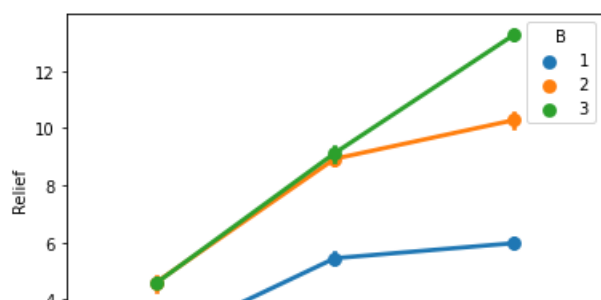
What is the interaction between the two treatments? [hint: use the 'pointplot' function from the 'seaborn' function]***
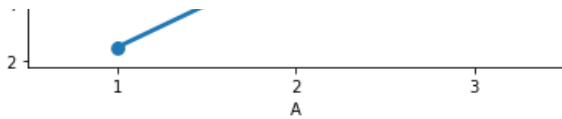
```
Step 1 - Using point plot and differentiating the variables with hue function
```

```
<AxesSubplot:xlabel='A', ylabel='Relief'>
```

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

From this output we can see that 1.Amount of active ingredients explain a significant amount of variation in average relief (p-values < 0.05). Hence reject the null hypothesis. This means that with the observed data, there is enough evidence to assume a general difference in the relief time of the active ingredients(1,2,3) when there is influence of A on B

Hence it shows that there is an interaction between the two ingredients

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) 2.0 220.020 110.010000 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

From this output we can see that 1.Amount of active ingredients explain a significant amount of variation in average relief (p-values < 0.05). Hence reject the null hypothesis. This means that with the observed data, there is enough evidence to assume a general difference in the relief time of the active ingredients(1,2,3) when there is influence of A on B

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

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

1. ANOVA test in this case study helped to compare A and B group

1. To determine whether a relationship exists between them and also if there is any variabilty between the groups

formula = 'Relief ~ C(A) + C(B) + C(A):C(B)' model = ols(formula, DF).fit() aov_table = anova_lm(model) print(aov_table)