A botanist wants to study the effects of sunlight (low vs. high) and watering frequency (daily vs. weekly) on the growth of a certain plant species.

2x2 factorial design will be an appropriate choice for studying the effect of sunlight and watering frequency on the growth of plant species. It will allow us to estimate both main effects and interaction effects between the two factors. So, we can determine independent effects of sunlight and watering frequency on plant growth as well as their combined effect on the growth of plant species. Also using factorial design will give a more precise estimation of treatment effects with smaller sample size.

In this case we have two factors involved:

- Sunlight (low vs high)
- Watering frequency (daily vs weekly)

The experimental schedule should be randomized to minimize the effects of potential confounding variables and ensure that the effects observed are due to the factors of interest and not due to other factors that were not controlled for. Randomization helps to ensure that each treatment combination has an equal chance of being assigned to any experimental unit, and thus reduces the potential bias in the estimates of the treatment effects.

Randomization would help to ensure that any variability observed in the response variable (plant growth) is due to the manipulated factors (sunlight and watering frequency) rather than other extraneous factors. For example, if the botanist placed all plants receiving high sunlight on one side of the garden and those receiving low sunlight on the other side, and all plants receiving daily watering in one section and weekly watering in another section, the potential for confounding effects such as soil nutrient differences, variations in temperature or humidity,

and insect or disease pressure could affect the results. By randomizing the experimental units (in this case, the plants), the botanist can ensure that any extraneous variables are distributed randomly among the treatment groups, minimizing the potential for bias in the results. It will also help to improve the precision and accuracy of the statistical analysis and increases the generalizability of the results to our population.

In this case the system output is plant growth, which is the response variable being measured. There are a total of 32 experiments conducted using a 2x2 factorial design with 2 replicates for each experiment. Each experiment is repeated twice, resulting in 2 observations for each of the 32 experiments.

In order to determine which of the main effects are active and whether the interactive effect is active,we can use both the regression table and the ANOVA table.

The regression table will provide information about the coefficients of each variable (sunlight, watering frequency), including the main effects and the interaction effect. If a coefficient has a significant p-value, typically set at 0.05 or lower, then the corresponding effect is considered active.

The ANOVA table will provide information about the overall significance of each effect. If the p-value associated with an effect is significant, then that effect is considered active. Additionally, it provides information about the relative importance of each effect, as measured by the F-value.

Regression Table

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.6278     0.6455  10.267  7.8e-15 ***
sunlight+          2.5467     0.9129   2.790  0.00706 **
water+            -0.4152     0.9129  -0.455  0.65087
sunlight+:water+  -1.6874     1.2911  -1.307  0.19620
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.582 on 60 degrees of freedom
Multiple R-squared:  0.1721,    Adjusted R-squared:  0.1307
F-statistic: 4.157 on 3 and 60 DF,  p-value: 0.009646
```

Interpretation:

The term "sunlight+" in the coefficients table represents the main effect of sunlight, while "water+" represents the main effect of water. The coefficient for "sunlight+" is 2.5467, which is positive and statistically significant (Pr(>|t|) = 0.00706), indicating that increasing the level of sunlight from low to high results in an increase in the response variable (growth). Similarly, the coefficient for "water+" is negative, but not statistically significant (Pr(>|t|) = 0.65087), suggesting that changing the level of water does not have a significant effect on growth.

The term "sunlight+: water+" represents the interactive effect between sunlight and water. The coefficient for this term is -1.6874, which is negative, but not statistically significant (Pr(>|t|) = 0.19620), indicating that the effect of changing both sunlight and water together is not significant in predicting growth.

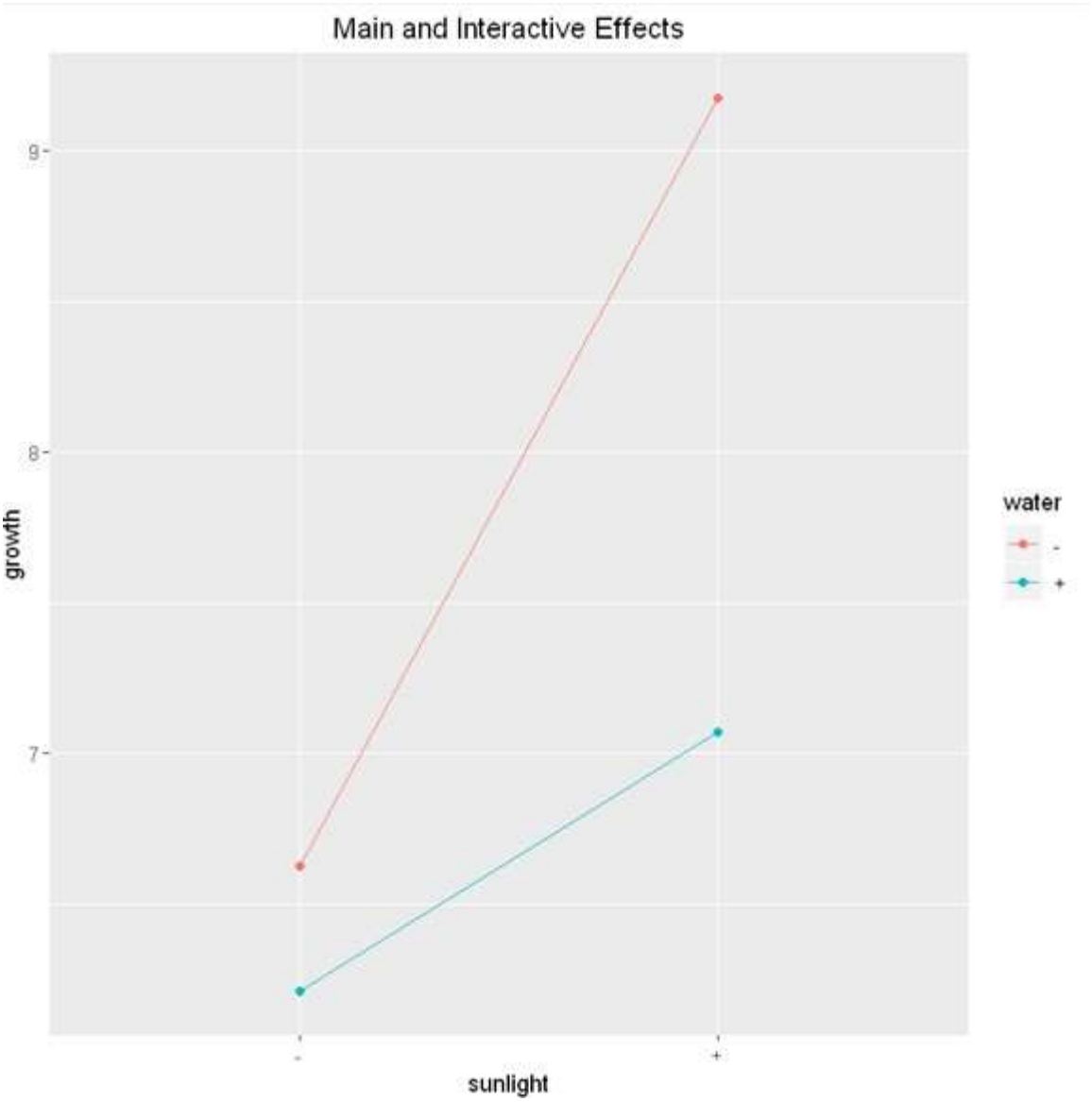
ANOVA table:

```
                Df Sum Sq Mean Sq F value Pr(>F)
sunlight         1   46.4   46.40   6.959 0.0106 *
water            1   25.4   25.36   3.803 0.0558 .
sunlight:water   1   11.4   11.39   1.708 0.1962
Residuals       60  400.0    6.67
---
```

The ANOVA table provides information on the significance of each main effect and the interactive effect. The F-statistic for the main effect of sunlight is 6.959 with a p-value of 0.0106, indicating that the main effect of sunlight is significant in predicting growth. The F-statistic for the main effect of water is 3.803 with a p-value of 0.0558, suggesting that the main effect of water is not significant. The F-statistic for the interactive effect is 1.708 with a p-value of 0.1962, which is not significant, indicating that the interactive effect is not significant.

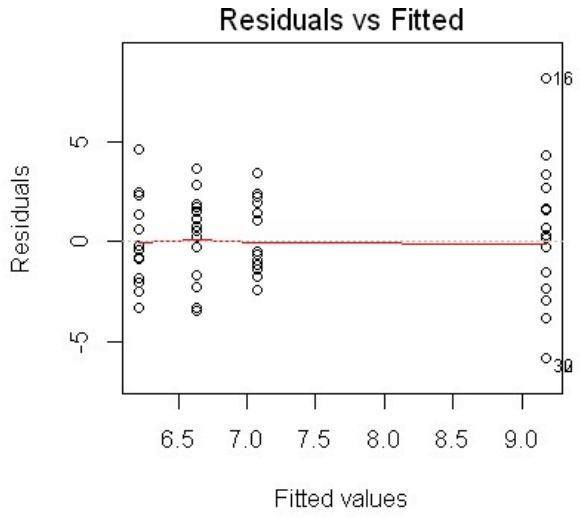
Therefore, based on the coefficients and the ANOVA table, we can conclude that the main effect of sunlight is active, and the main effect of water is not active, and the interactive effect between sunlight and water is not active.

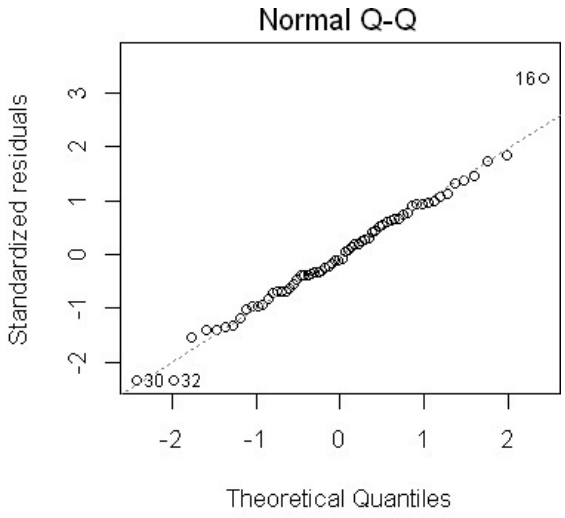
To determine the best combination of sunlight and water. We can use the plot of Main and Interactive Effects. In this plot, we will find the best combination of sunlight and water by finding the point where the response variable (y-axis) is maxim

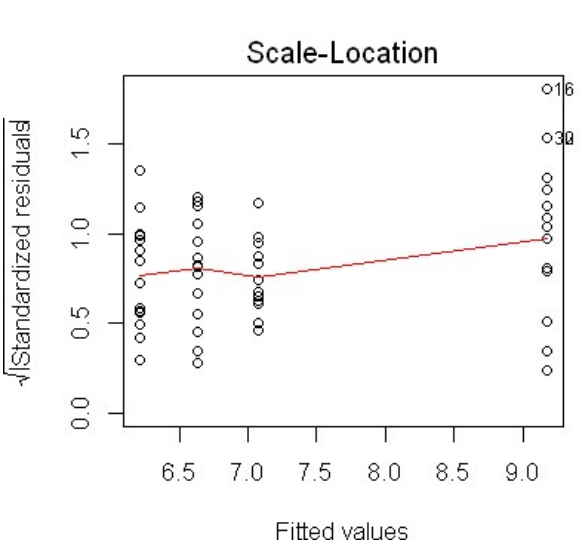


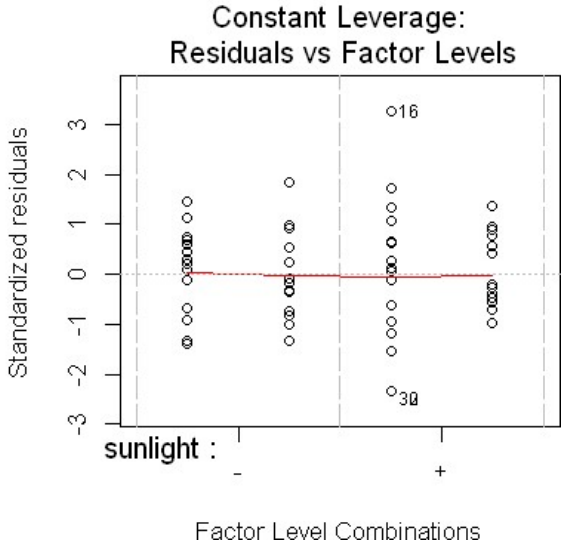
From the plot, it appears that the highest response is obtained at the combination of high sunlight and low water, indicating that this is the best combination for maximizing the response variable.

b. Consider the four residual distribution plots, the Shapiro-Wilk test and the Lilliefors test.Are the residuals normally distributed?


Residuals vs Fitted
Residuals
Fitted values
16
30


Normal Q-Q
Standardized residuals
Theoretical Quantiles
16
30
32


Scale-Location
√|Standardized residuals|
Fitted values
16
30


Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
sunlight :
Factor Level Combinations
16
30

To determine if the residuals are normally distributed, we can visually inspect the residual distribution created. The plots are:

Shapiro-Wilk test

Based on the data p-value is greater than the significance level (0.05), hence, we conclude that the data is normally distributed.

Normal Q-Q plot:

We can see a linear line which is not completely straight but also not badly skewed. Hence, we can conclude that the residuals are normally distributed.

To summarize, we concluded that the residuals are normally distributed by inspecting the four plots.

To ensure validity of the obtained and interpreted results we need to verify that following assumptions have been met:

1. Normality of Residuals: The residuals should be normally distributed. (already done)
2. Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variables.
3. Independence: The residuals should be independent of each other.
4. Linearity: This assumes that the relationship between the dependent and independent variables is linear. This can be verified by plotting the dependent variable against each independent variable and ensuring that the relationship is roughly linear.
5. Homoscedasticity: This assumes that the variance of the errors is constant across all levels of theindependent variables. This can be verified by plotting the residuals against the predicted valuesand ensuring that the spread of the residuals is roughly constant across all levels of the predicted values.
6. Independence: This assumes that the errors are independent of each other. This can be verified by plotting the residuals against and ensuring that there are no systematic patterns in the residuals over time.

To check these assumptions, we can develop the following three plots:

1. Q-Q Plot: This plot is used to verify normality of residuals. It plots the quantiles of the residuals against the quantiles of a normal distribution. If the residuals follow a straight line, then they are normally distributed.

2. Residuals vs. Fitted Values Plot: This plot is used to check for homoscedasticity. It plots the residuals against the fitted values (predicted values). If the variance of the residuals is consistent across all levels of the fitted values, then homoscedasticity is met.
3. Autocorrelation Plot: This plot is used to check for independence of residuals. It plots the residuals against their lagged values (i.e., the residuals from the previous time step). If the residuals are not correlated with their lagged values, then independence is met.