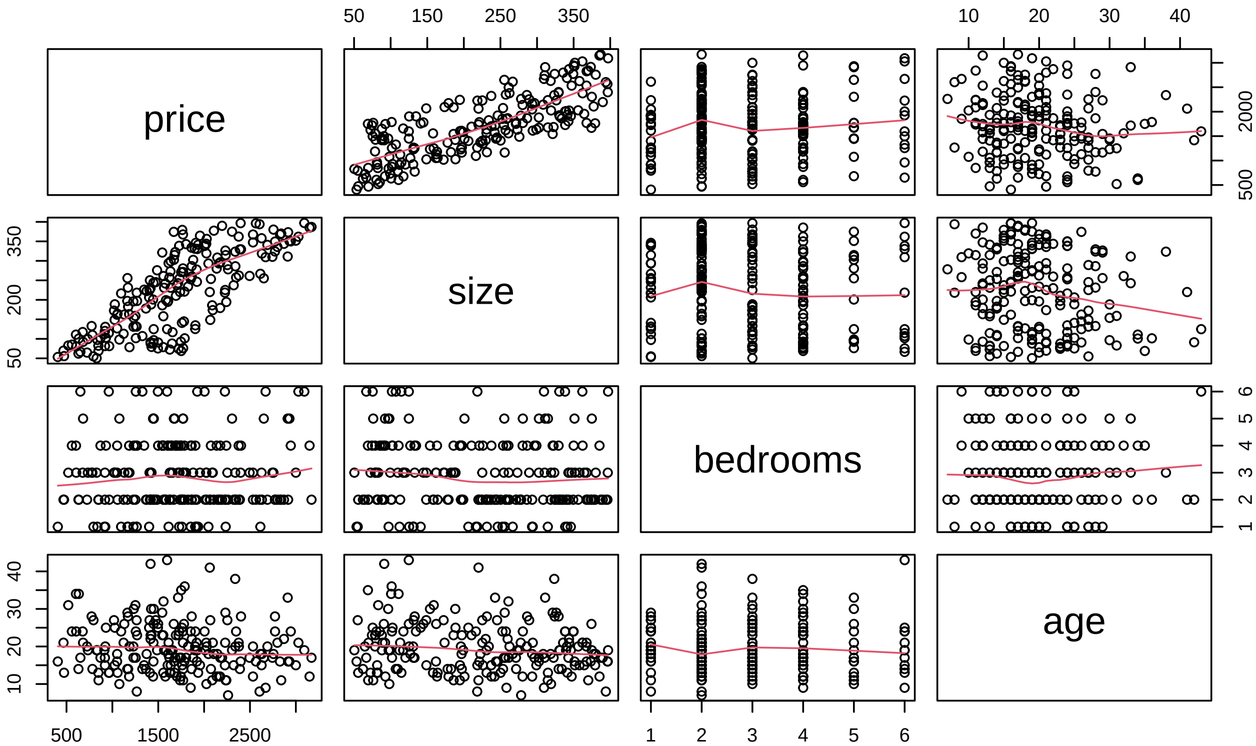
PROJECT REPORT

Nhi Doan

Question 1

a. Inputting the data and producing the **scatterplot** and **correlation matrix**:

```
realest = read.csv('realestate2024.csv', header = T)
pairs(realest, panel = panel.smooth)
```




```
cor(realest)
```

```
##               price        size    bedrooms         age
## price     1.00000000  0.77994644  0.05560245 -0.12347514
## size      0.77994644  1.00000000 -0.07285563 -0.16695401
## bedrooms  0.05560245 -0.07285563  1.00000000  0.02850195
## age      -0.12347514 -0.16695401  0.02850195  1.00000000
```

- The response variable price has a strong positive linear relationship with the predictor `size`; a weak negative linear relationship with predictor `age` and no obvious correlation with the predictor `bedrooms`.
- There seems to be a very weak negative correlation between predictor `size` and predictor `age`. In overall, there is no obvious relationship among the predictors themselves.

b. Fit the **full model**:

```
realest.lm = lm(price ~ size + bedrooms + age, data = realest)
summary(realest.lm)
```

```
##
## Call:
## lm(formula = price ~ size + bedrooms + age, data = realest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -748.08 -318.57  -54.74  366.46  784.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 449.2955   133.7219   3.360  0.00094 ***
## size          4.9371     0.2819  17.514  < 2e-16 ***
## bedrooms     53.6872    21.1222   2.542  0.01182 *
## age           0.4821     4.3038   0.112  0.91092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.7 on 193 degrees of freedom
## Multiple R-squared:  0.621,  Adjusted R-squared:  0.6152
## F-statistic: 105.4 on 3 and 193 DF,  p-value: < 2.2e-16
```

```
summary.realest = summary(realest.lm)
se.size=sqrt(diag(summary.realest$sigma^2 * summary.realest$cov.unscaled))[2]
```

The **required CI** is

$$
\begin{aligned}
&\hat{\beta}_{\text{size}} \pm t_{n-p,1-\alpha/2}\,\text{s.e.}(\hat{\beta}_{\text{size}}) \\
&= \hat{\beta}_{\text{size}} \pm t_{193,0.975}\,\text{s.e.}(\hat{\beta}_{\text{size}}) \\
&= 4.9371 \pm 1.972332 \times 0.2818869 \\
&= (4.381125, 5.493075)
\end{aligned}
$$

We are 95% confident that for every square meter increase in size, the price will increase on average between \$4,381.125 and \$5,493.075.

c. **Regression Model**:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, 2, ...n$$

– Y: the response variable `price`;
– X_{ij}: the predictors variables for the j-th observation
$*X_{i1}$: The `size` of the property (in square meters)

*X_{i2}: The number of `bedrooms` in the property
*X_{i3}: The `age` of the property in years
– β_0: the intercept of the regression model;
– β_1, β_2, β_3: the coefficients of the predictors variables `size`, `bedrooms`, `age` respectively;
– $\epsilon \sim N(0, \sigma^2)$: denotes the random variation with constant variance;

Conducting the F-test, we have,

- **Hypotheses:**
  - $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
  - $H_1 :$ Not all $\beta_i \neq 0, \quad i = 1, 2, 3$
- Standard R output **ANOVA table**:

```
anova(realest.lm)
```

```
## Analysis of Variance Table
##
## Response: price
##            Df   Sum Sq  Mean Sq  F value  Pr(>F)
## size        1 49256631 49256631 309.8153 < 2e-16 ***
## bedrooms    1  1028915  1028915   6.4717 0.01174 *
## age         1     1995     1995   0.0125 0.91092
## Residuals 193 30684511   158987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

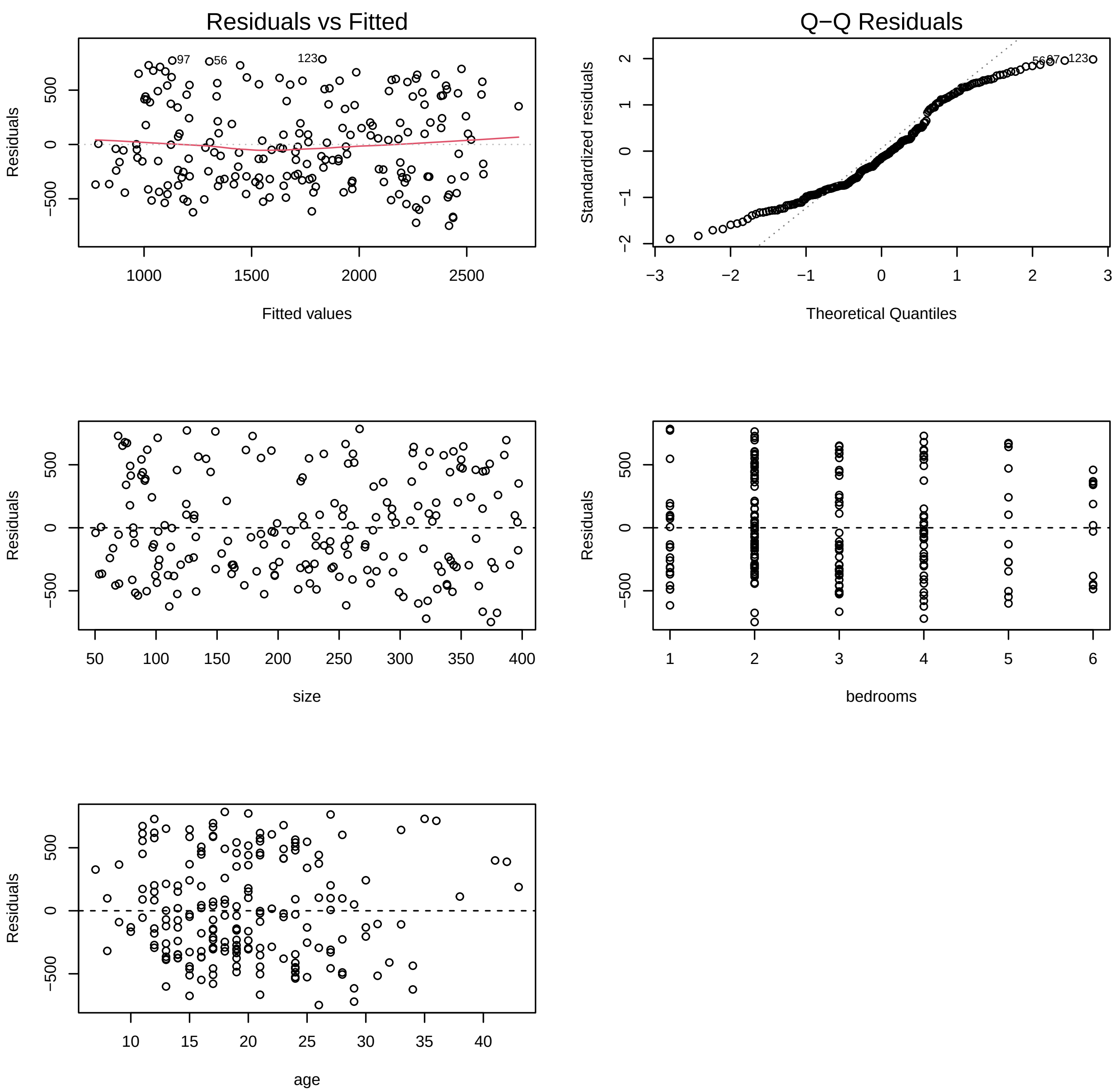
- The reduced Overall ANOVA table:

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Regression	3	50287541	16762514	105.43323	0
Residuals	193	30684511	158987		

- Note the Reg SS = 49256631 + 1028915 + 1995 = 50287541
- Therefore the $MS_{Reg} = SS_{Reg}/df_{Reg} = 50287541/3 = 16762514$
- Test statistic: $F_{obs} = MS_{Reg}/MS_{Res} = 16762514/158987 = 105.4332$
- The null distribution for the test statistics is F(3,193)
- P-value: P $(F_{3,193} \geq 105.4332) = 0 = 1.289351\text{e-}24 < 0.05$
- **Conclusion**: As the P-value is very small,
  - (Statistical) There is enough evidence to reject H_0.
  - (Contextual) There is a significant linear relationship between `price` and at least one of the 3 predictors variables.

d. For the diagnostics:

```r
par(mfrow = c(3, 2))
plot(realest.lm, which = 1:2)
plot(resid(realest.lm) ~ size, data = realest, xlab = "size", ylab = "Residuals")
abline(h = 0, lty = 2)
plot(resid(realest.lm) ~ bedrooms, data = realest, xlab = "bedrooms", ylab = "Residuals")
abline(h = 0, lty = 2)
plot(resid(realest.lm) ~ age, data = realest, xlab = "age", ylab = "Residuals")
abline(h = 0, lty = 2)
```



- The quantile plot of the residuals look approximately linear, which means the residuals are normally distributed. This supports the assumption that our model's residuals follow a normal pattern.

- The residual plots do not show any clear patterns, indicating that the assumptions of linearity and constant variance are also met.

Overall, this suggests that our multiple linear regression model is appropriate for explaining property prices.

e. We have $\mathbf{R^2 = 0.621 = 62.1\%}$, meaning that 62.1% of the variation in property `price` can be explained by the full linear regression model. This indicates that the model performs a reasonably good job of capturing the variables influencing property `price`.

f. Starting with all the predictors

```r
summary(realest.lm)
```

```
##
## Call:
## lm(formula = price ~ size + bedrooms + age, data = realest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -748.08 -318.57  -54.74  366.46  784.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 449.2955   133.7219   3.360  0.00094 ***
## size          4.9371     0.2819  17.514  < 2e-16 ***
## bedrooms     53.6872    21.1222   2.542  0.01182 *
## age           0.4821     4.3038   0.112  0.91092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.7 on 193 degrees of freedom
## Multiple R-squared:  0.621,  Adjusted R-squared:  0.6152
## F-statistic: 105.4 on 3 and 193 DF,  p-value: < 2.2e-16
```

- `age` has the highest P-value(0.91092) so we shall remove it first.

```r
realest.lm2 = lm(price ~ size + bedrooms, data=realest)
summary(realest.lm2)
```

```
##
## Call:
## lm(formula = price ~ size + bedrooms, data = realest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -744.25 -321.86  -59.73  362.39  783.79
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 459.8715    94.4581   4.869 2.33e-06 ***
## size          4.9318     0.2773  17.785  < 2e-16 ***
## bedrooms     53.7265    21.0655   2.550   0.0115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 397.7 on 194 degrees of freedom
## Multiple R-squared:  0.621,  Adjusted R-squared:  0.6171
## F-statistic:   159 on 2 and 194 DF,  p-value: < 2.2e-16
```

- At this point, all remaining predictors are significant and should be retained in the model. The final fitted model equation is

$$\hat{Y} = 459.8715 + 4.9318X_1 + 53.7265X_2$$

$$\text{or} \quad \texttt{price} = 459.8715 + 4.9318 \times \texttt{size} + 53.7265 \times \texttt{bedrooms}$$

f. The R^2 indicates how much of the variation in property prices is explained by the predictors in the model. In both the full model and the final model, the R-square remains consistent at approximately 0.621, suggesting that the model explains about 62.1% of the variation in property prices, regardless of whether predictor age is included in the model.

   However, the adjusted R^2 increases slightly from 61.52% in the full model to 61.71% in the final model. This increase occurs because the adjusted R-square penalizes for the number of predictors, meaning it can improve when non-significant predictors are removed. Thus, while the R^2 did not change, the increase in adjusted R^2 indicates that the final model is a better parsimonious model for the data.

Question 2

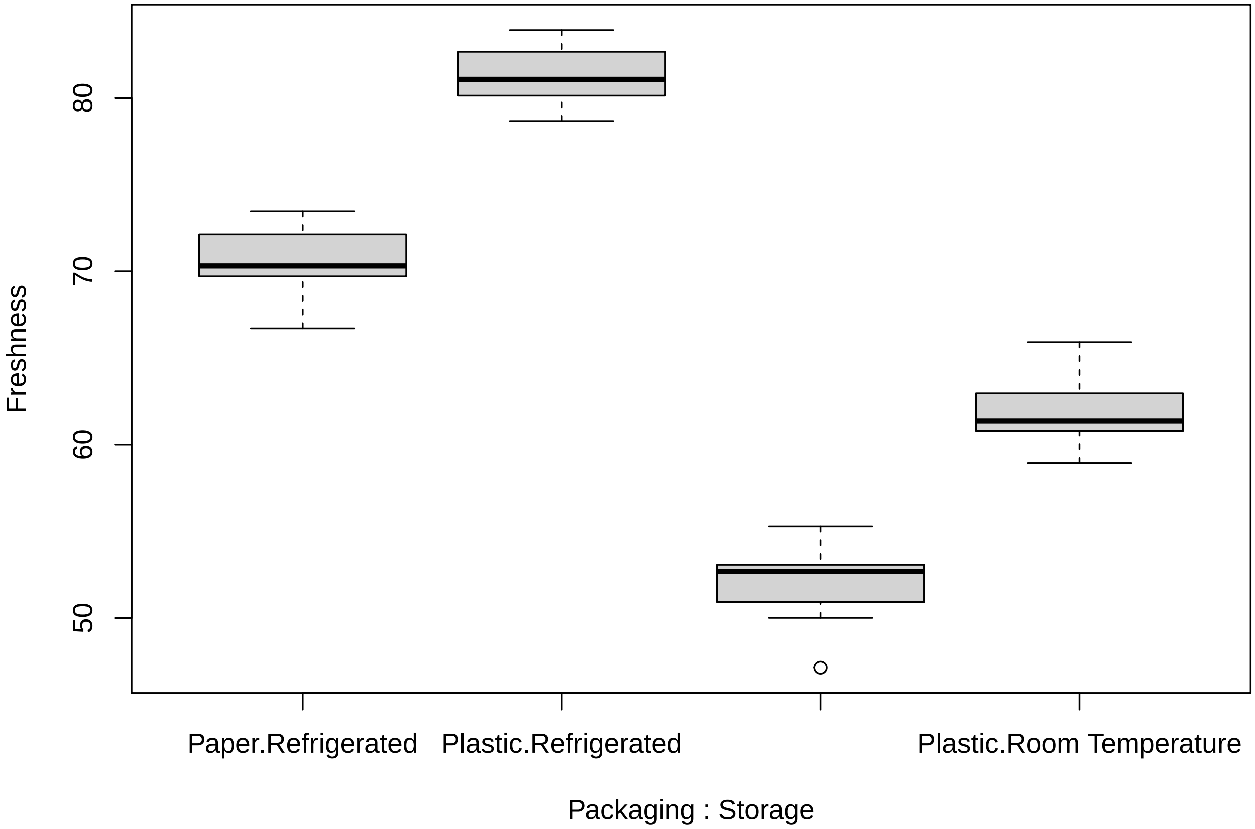
a. A study is considered balance if each combination of factors has the same number of replicates. For this study, we have:

```r
goods = read.csv('goods.csv', header=T, stringsAsFactors = T)
table(goods[, c("Packaging", "Storage")])
```

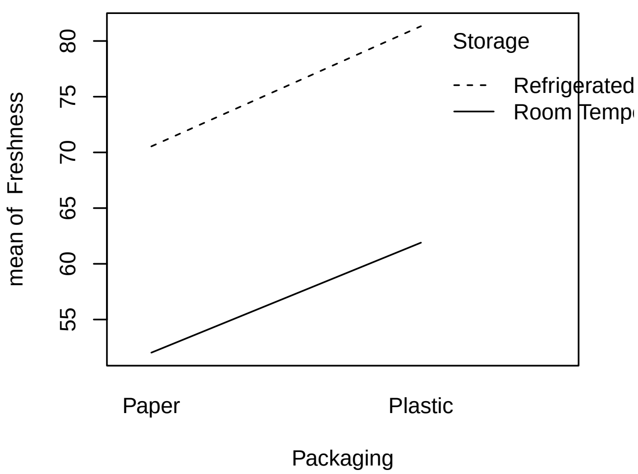
```
##          Storage
## Packaging Refrigerated Room Temperature
##   Paper             14               17
##   Plastic           16               18
```

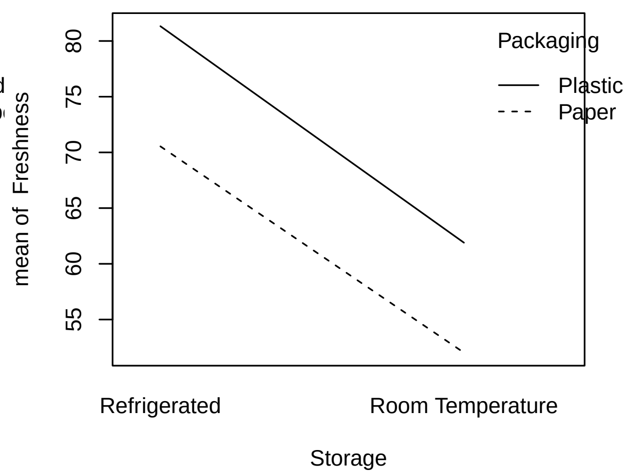
- We can see that the design is unbalanced with an unequal number of replicates for each combination of levels of the two factors.

```r
boxplot(Freshness ~ Packaging + Storage, data = goods)
```




```r
par(mfrow = c(2, 2))
with(goods, interaction.plot(Packaging, Storage, Freshness))
with(goods, interaction.plot(Storage, Packaging, Freshness))
```





- From the boxplot, it appears that the assumption of equal variance among the different levels of the factors is approximately valid, as indicated by the similar sizes of the boxes.

- The interaction plots indicate that the lines are parallel, suggesting that there is no significant interaction between `Packaging` type and `Storage` condition in influencing the `Freshness` score. This means that the effect of `Storage` condition on `Freshness` is consistent across different `Packaging` types, and vice versa.

c. Two-Way ANOVA model with interaction:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

- Y_{ijk}: the `Freshness` score response;
- α_i: the `Packaging` effect, there are two levels - Plastic and Paper
- β_j: the `Storage` effect, there are two levels - Room Temperature and Refrigerated
- γ_{ij}: interaction effect between `Packaging` and `Storage`
- ϵ_{ijk} ~ N(0,σ^2): the unexplained variation, normally distributed.

- Since the interaction effect is not significant (as indicated by the parallel lines), the appropriate model is a **Two-Way ANOVA model without interaction**:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

d. We have found that the interaction effect is not significant, therefore, it is appropriate to exclude the interaction term in our Two-Way ANOVA model. However, we wish to test the full model:

Hypotheses:

- Interaction effect: H_0 :$\gamma_{ij} = 0$ for all i,j vs H_1 : at least one $\gamma_{ij} \neq 0$
- Main effect of Packaging: H_0: $\alpha_i = 0$ for all i vs H_1: at lease one $\alpha_i \neq 0$
- Main effect of Storage: H_0: $\beta_j = 0$ for all j vs H_1: at lease one $\beta_j \neq 0$

Fitting this interaction model:

```
full.aov = lm(Freshness ~ Packaging * Storage, data = goods)
anova(full.aov)
```

```
## Analysis of Variance Table
##
## Response: Freshness
##                   Df Sum Sq Mean Sq   F value Pr(>F)
## Packaging          1 1839.8  1839.8  613.6776 <2e-16 ***
## Storage            1 5824.5  5824.5 1942.7752 <2e-16 ***
## Packaging:Storage  1    3.4     3.4    1.1295 0.2921
## Residuals         61  182.9     3.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We can see that the interaction terms are not significant, since it has a P-value of 0.2921 > 0.05. They can be removed from the model.

The model can be revised as:

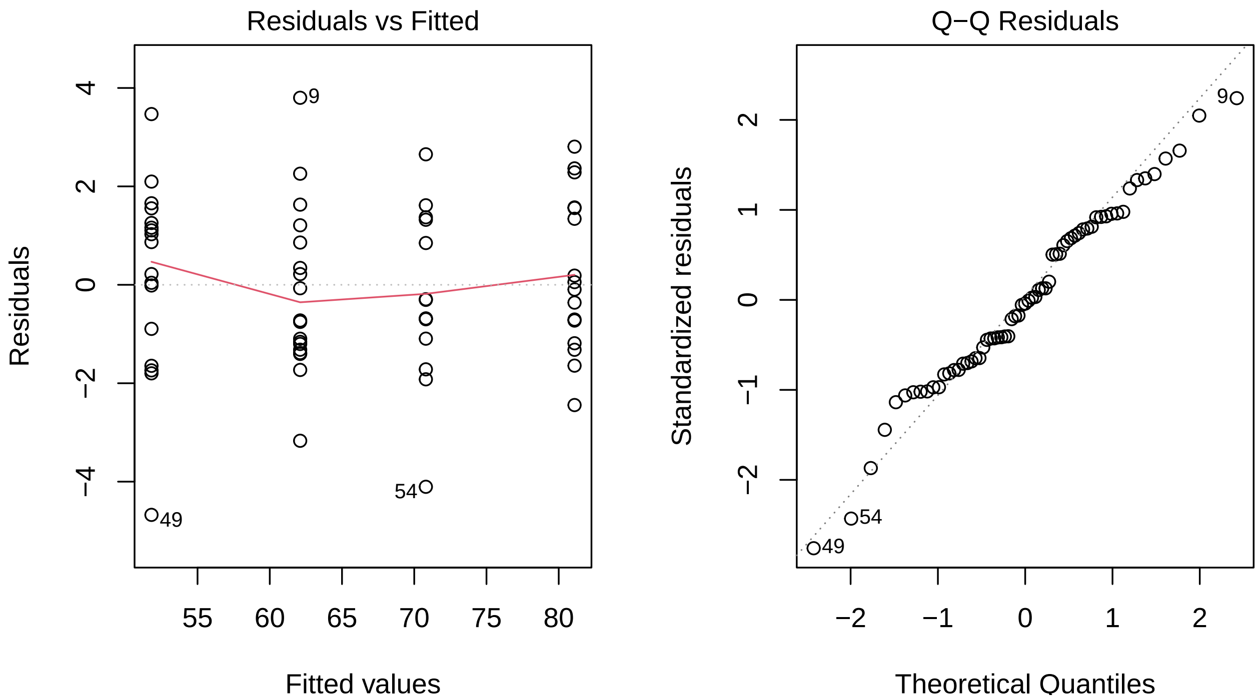
```
revised.aov = lm(Freshness ~ Packaging + Storage, data = goods)
anova(revised.aov)
```

```
## Analysis of Variance Table
##
## Response: Freshness
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Packaging  1 1839.8  1839.8   612.4 < 2.2e-16 ***
## Storage    1 5824.5  5824.5  1938.7 < 2.2e-16 ***
## Residuals 62  186.3     3.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We can see that both `Packaging` and `Storage` have significant effects on `Freshness` (P-value of 2.2e-16<0.05). Therefore, they can not be removed from the model. This means we reached our final model.

We check the assumptions of the revised model with the diagnostic plots:

```
par(mfrow =c(1,2))
plot(revised.aov, which = 1:2)
```



- Most points follow the diagonal line, suggesting that the residuals follow a normal distribution.

The residuals appear randomly scattered around the fitted values, with no obvious trend or pattern, indicating that the constant variance assumption is satisfied.

- Based on the diagnostic plots, the assumptions for the model without interaction are met. In conclusion, the **best model is Two-Way ANOVA model without interaction**.