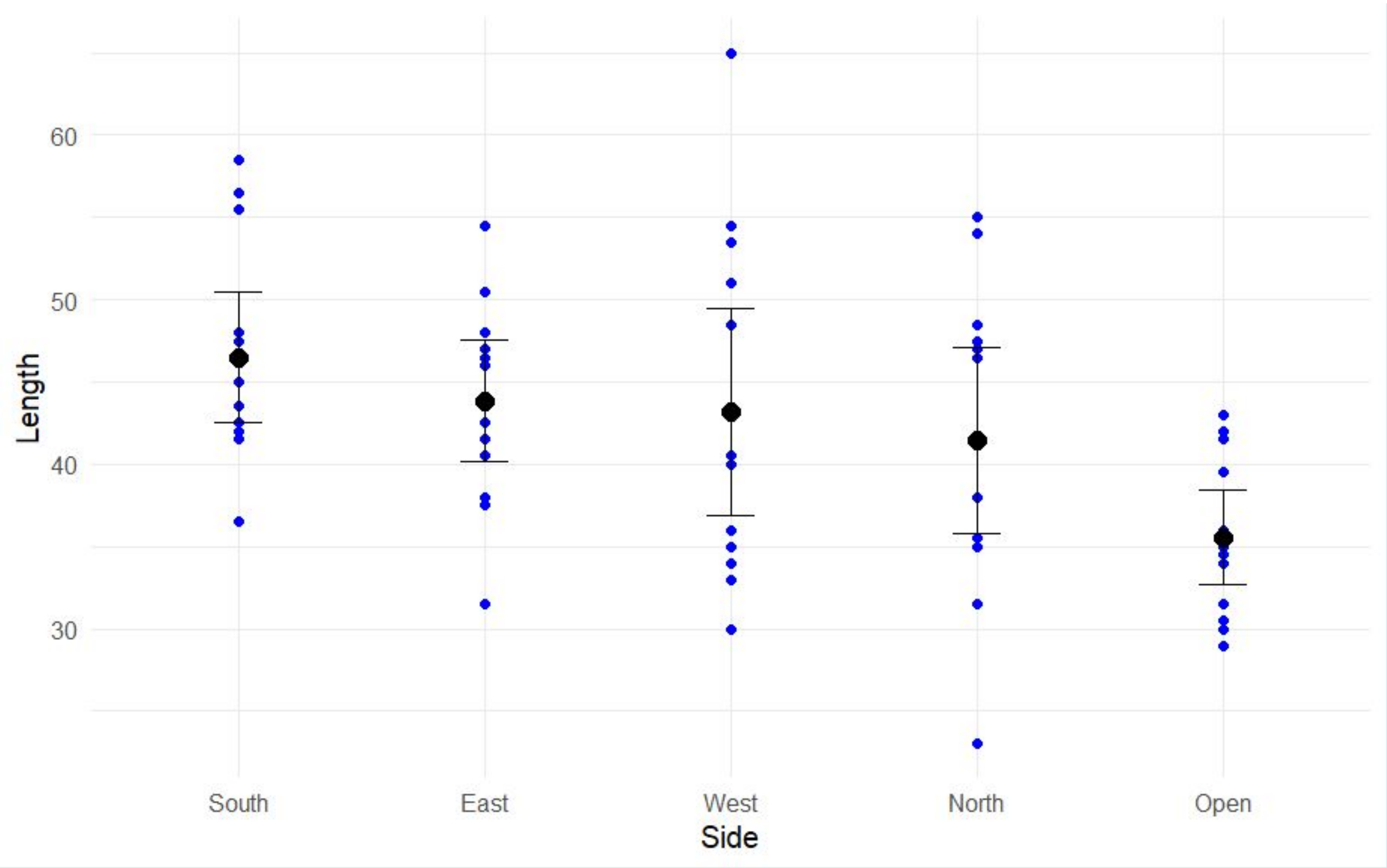
Exercise 7

Laura Fernández, Kristin Olsen and Lea Frei

Code for making the figure:

```
daffodil_data <- read.table("data/daffodils.csv", header = T, sep = ",")

summary_data <- daffodil_data %>%
  group_by(Side) %>%
  summarize(
        Mean = mean(Length),
        Lower_CI = Mean - 1.96 * sd(Length) / sqrt(n()),
        Upper_CI = Mean + 1.96 * sd(Length) / sqrt(n()))

side.ord <- factor(daffodil_data$Side, c("South", "East", "West", "North", "Open"))

ggplot(daffodil_data, aes(x = side.ord, y = Length), ylim = c(20, 90)) +
  geom_point(colour = "blue") +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar",
        width = 0.2) +
  stat_summary(fun = mean, geom = "point", size = 3) +
  theme_minimal() +
  labs(x = "Side", y = "Length")
```



60
50
40
30
Length
South
East
West
North
Open
Side

A researcher collected daffodils (flowers) from four sides of a building and from an open area nearby. She wondered whether the **average stem length** of a daffodil **depends on its location**. The data set is available as daffodils.csv from ILIAS.

1. Exploration of the data

```
daffodil_data <- read.table("data/daffodils.csv",
                            header = T, sep = ",")
```

length : response variable

side: categorical variable (explanatory)

5 labels: East, West, North, South and Open Side

	Length	Side
9	46.5	East
10	37.5	East
11	31.5	East
12	54.5	East
13	50.5	East
14	33.0	West
15	30.0	West
16	35.0	West
17	36.0	West
18	40.5	West
19	53.5	West
20	51.0	West
21	65.0	West

a) State the null hypothesis of an ANOVA model for this problem.

Brief reminder of ANOVA

Compare the variance **within** groups to variance **between** groups.

If the variance between groups is significantly larger than the variance within groups, it suggests that at least one group mean is different from the others.

Basic assumption: variance (“amount of randomness”) is the same in each group.

- Often used for categorical explanatory variables
- It is possible to fit the ANOVA model with multiple linear regression.

a) State the null hypothesis of an ANOVA model for this problem

In words:

5 groups of flowers based on the location: East, North, Open Side, South, West

Null hypothesis : There is **no significant difference** in the average stem length between the different groups.

Alternative hypothesis: There is significant difference in the average stem length of the flowers between the different groups.

a) State the null hypothesis of an ANOVA model for this problem

As a formula :
Given the model:

$$Y_{ij} = \mu + \alpha_i + E_{ij}$$

Cell means model

α_i : group effect for i = 2 , …, 5,
where 1 = East, 2 = North, 3 = Open, 4 = South , 5 = West,
and $E_{ij} \sim^{\text{i.i.d.}} N(0, \sigma^2)$ for i = 2, … , 5, j = 1, …, n_i.

The null hypothesis is:

$$H_0: \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$$

Why do we not include α_1 ?

Why do we not include α_1 ?

$$Y_{ij} = \mu + \alpha_i + E_{ij}, \quad E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

The model is overparameterized.
We assume the condition:

$$\alpha_1 = 0$$

We must fit 5 parameters: μ , α_2 , α_3, α_4 , α_5

- **Fit into multiple linear regression model**

Transform categorical variables to artificial explanatory variables:

$$x_1 \begin{cases} 1 \text{ , when the daffodil belongs to North} \\ 0 \text{ , otherwise} \end{cases}$$

Fit into multiple linear regression model

Transform categorical variables to artificial explanatory variables:

$$x_1 \begin{cases} 1\text{ , when the daffodil belongs to North} \\ 0\text{ , otherwise} \end{cases}$$

$$x_2 \begin{cases} 1\text{ , when the daffodil belongs to Open Side} \\ 0\text{ , otherwise} \end{cases}$$

$$x_3 \begin{cases} 1\text{ , when the daffodil belongs to South} \\ 0\text{ , otherwise} \end{cases}$$

$$x_4 \begin{cases} 1\text{, when the daffodil belongs to West} \\ 0\text{ , otherwise} \end{cases}$$

$$\boxed{Y_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + E_j}$$

Fit into multiple linear regression model

Cell means model

$$Y_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + E_j$$

$$Y_{ij} = \mu + \alpha_i + E_{ij} \,,$$

Correspondence of coefficients:

$\beta_0 = \mu$

$\beta_1 = \alpha_2$

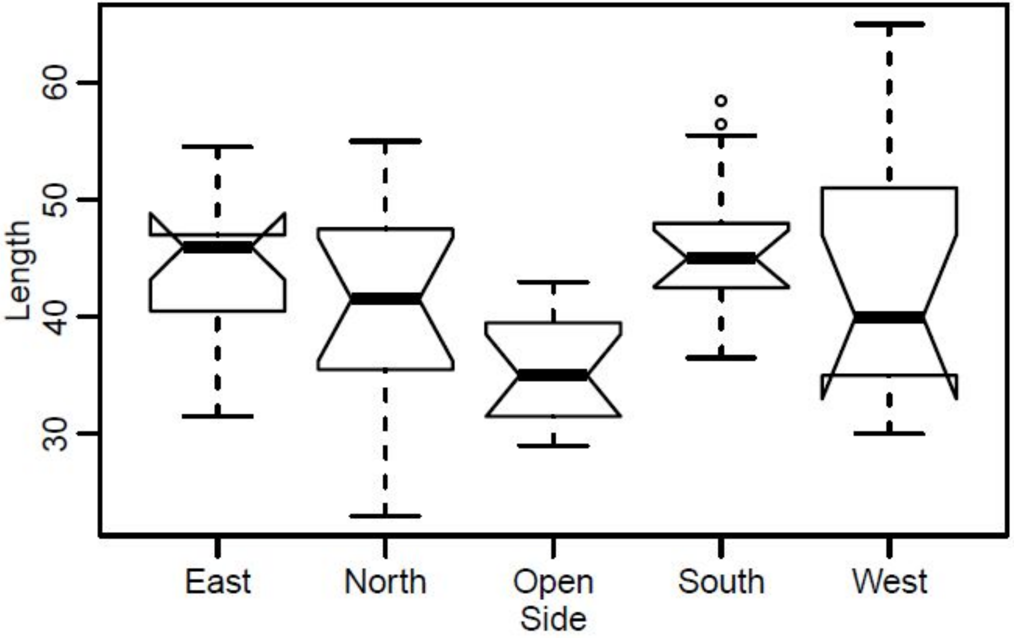
$\beta_2 = \alpha_3$

$\beta_3 = \alpha_4$

$\beta_4 = \alpha_5$

α_i : group effect for i = 2 , …, 5, where 1 = East, 2 = North, 3 = Open, 4 = South , 5 = West and $E_{ij} \sim^{i.i.d.} N(0, \sigma^2)$ for i = 2, … , 5, j = 1, …, n_i.

b) A boxplot of the data looks as follows:
Does it appear that the null hypothesis is true?



- The variance within all the groups seems to be similar, except the **Open Side group**.
- The range of values in the Open Side group is narrower.
- The effect, α_3 might be different from 0.
- The null hypothesis appears to be false.

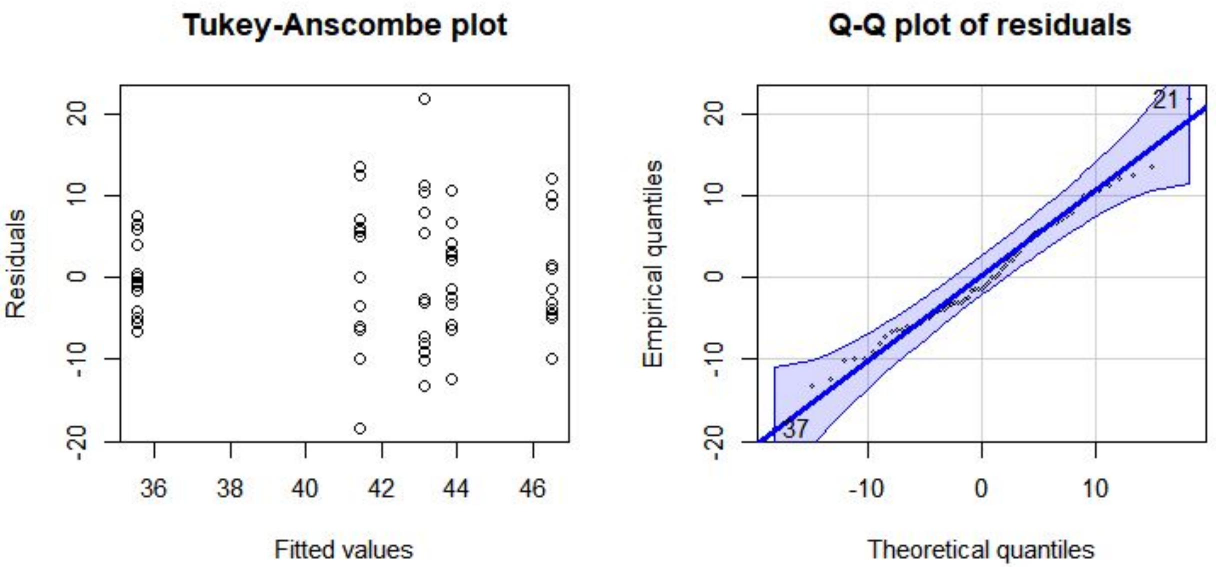
c) Fit an ANOVA model to the data and test the null hypothesis from a) on a significance level of 10%.

```
> daffodils.fit <- lm(Length ~ Side, data = daffodil_data)
> summary(daffodils.fit)

Call:
lm(formula = Length ~ Side, data = daffodil_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.423  -5.038  -1.346   5.577  21.846 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  43.8462     2.1449  20.442  < 2e-16 ***
SideNorth    -2.4231     3.0334  -0.799  0.42755    
SideOpen     -8.3077     3.0334  -2.739  0.00811 ** 
SideSouth     2.6538     3.0334   0.875  0.38513    
SideWest     -0.6923     3.0334  -0.228  0.82024    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.734 on 60 degrees of freedom
Multiple R-squared:  0.1954,	Adjusted R-squared:  0.1417 
F-statistic: 3.642 on 4 and 60 DF,  p-value: 0.01009
```

c) Fit an ANOVA model to the data and test the null hypothesis from a) on a significance level of 10%.

```
> anova(daffodils.fit)
Analysis of Variance Table

Response: Length
          Df Sum Sq Mean Sq F value  Pr(>F)
Side       4  871.4 217.852  3.6425 0.01009 *
Residuals 60 3588.5  59.809
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) Does the ANOVA model fit well to the data? Perform a residual analysis




```
par(mfrow = c(1,2), cex = 0.5)
plot(fitted(daffodils.fit), resid(daffodils.fit),
     xlab = "Fitted values", ylab = "Residuals", main = "Tukey-Anscombe plot")
qqPlot(resid(daffodils.fit), dist = "norm",
       mean = mean(resid(daffodils.fit)),
       sd = sd(resid(daffodils.fit)),
       xlab = "Theoretical quantiles", ylab = "Empirical quantiles",
       main = "Q-Q plot of residuals")
```

e)

Which locations (sides of the building and open area) are not significantly different on a 5% level? Use Bonferroni adjusted pairwise t-tests.

e)

Which locations (sides of the building and open area) are not significantly different on a 5% level? Use Bonferroni adjusted pairwise t-tests.

ANOVA tests the null hypothesis that all the means do not differ from each other.

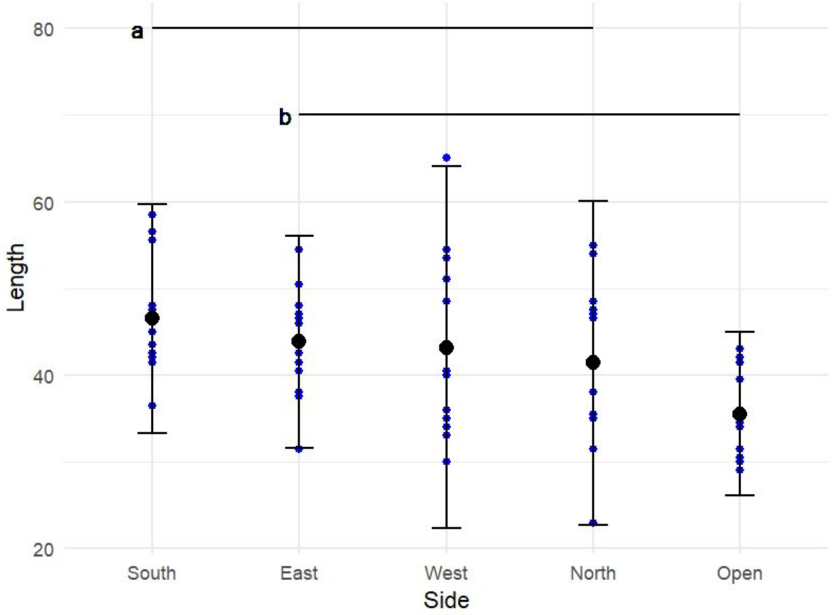
Pairwise t-tests to find which groups' means differ

e)

```
pairwise.t.test(daffodil_data$Length, daffodil_data$Side,
                p.adjust.method = "bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  daffodil_data$Length and daffodil_data$Side

      East   North  Open   South
North 1.0000 -      -      -
Open  0.0811 0.5709 -      -
South 1.0000 0.9940 0.0062 -
West  1.0000 1.0000 0.1477 1.0000

P value adjustment method: bonferroni
```

Only the means of open area and south side differ significantly on a 5% level e)

```
pairwise.t.test(daffodil_data$Length, daffodil_data$Side,
                p.adjust.method = "bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  daffodil_data$Length and daffodil_data$Side

      East   North  Open   South
North 1.0000 -      -      -
Open  0.0811 0.5709 -      -
South 1.0000 0.9940 0.0062 -
West  1.0000 1.0000 0.1477 1.0000

P value adjustment method: bonferroni
```



Only the means of open area and south side differ significantly on a 5% level

Conclusion

- example of how to fit an ANOVA model to data with several categorical explanatory variables
- tests the null hypothesis that none of the group means differ significantly
- pairwise t-tests to find out which groups actually differ in their means