Analysis of Orchid and Hospital Data

Will Harrison

```
library(dplyr, warn.conflicts = FALSE)
library(lubridate, warn.conflicts = FALSE)
library(tidyr)
library(ggplot2)
library(patchwork)
```

Question 1 [19 marks]

An orchid grower delivered a large sample of orchids to a distributor on 20 October 2022. Each orchid's height was recorded in inches and each orchid was assigned a score between 0 and 10 (0=very poor quality, 10=excellent quality). Any orchid with a score above 6 is bought by the distributor, while a score of 6 or lower leads to the orchid not being bought by the distributor.

The orchid grower asks you to analyze the data they collected. In addition to the height and score, you are given the type of orchid, the temperature at which the plant was grown, the levels of phosphate, potassium and sulfur levels used for fertilization, and the date the orchid was transferred to an individual pot in spring.

The full data are in the file "Orchids.csv" and a detailed data description is provided in the file "Data Descriptions.pdf".

a) *Load and clean the data. Extract and provide the first two rows of the data set. State the minimum and maximum observed phosphate, potassium and sulphur levels.* **[4 marks]**

```
Orchids <- read.csv("Orchids.csv")
```

After using **glimpse()** to look at the data (not shown), everything looks OK - just need to change the 'Planting' variable to the date data type.

```
Orchids$Planting <- ymd(Orchids$Planting)
head(Orchids, 2)
```

```
##   Height Phos Potas Sulf   Planting          Type Temp Quality
## 1   16.3   89   270   38 2022-03-19  Phalaenopsis 27.7       7
## 2    2.6    0   265   39 2022-04-01 Odontoglossum 18.1       5
```

Missing values for Phos, Pota and Sulf are indicated by a value of 0, therefore these will be discounted when finding the minimum.

```
Orchids %>%
  select(Phos, Potas, Sulf) %>%
  na_if(0) %>%
  summarise_all(list(min = min ,max = max), na.rm = TRUE)
```

```
##   Phos_min Potas_min Sulf_min Phos_max Potas_max Sulf_max
## 1       46       195       28      130       385       46
```

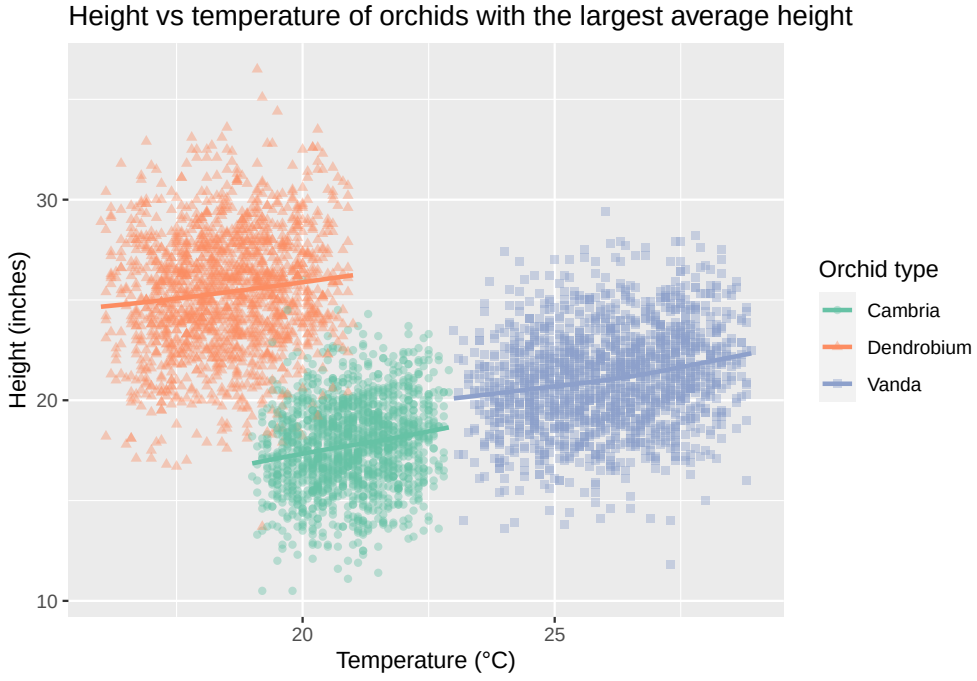
b) *Explore the relationship of temperature and plant height for the three types of orchid with the highest average height. Further investigate how these three types compare regarding their quality.* **[5 marks]**

```
Orchids %>%
  group_by(Type) %>%
  summarise("AvgHeight" = mean(Height)) %>%
  arrange(desc(AvgHeight)) %>%
  head(3)
```

```
## # A tibble: 3 x 2
##   Type       AvgHeight
##   <chr>          <dbl>
## 1 Dendrobium      25.4
## 2 Vanda           21.1
## 3 Cambria         17.8
```

```
Orchids_Filtered <- filter(Orchids, Type %in% c("Dendrobium", "Vanda",
                                                "Cambria"))

ggplot(Orchids_Filtered, aes(x = Temp, y = Height, colour = Type,
                             shape = Type)) +
  geom_point(alpha = 0.4) +
  scale_colour_brewer(palette = "Set2") +
  labs(title =
       "Height vs temperature of orchids with the largest average height",
       x = "Temperature (°C)", y = "Height (inches)", colour = "Orchid type",
       shape = "Orchid type") +
  geom_smooth(se = FALSE)
```



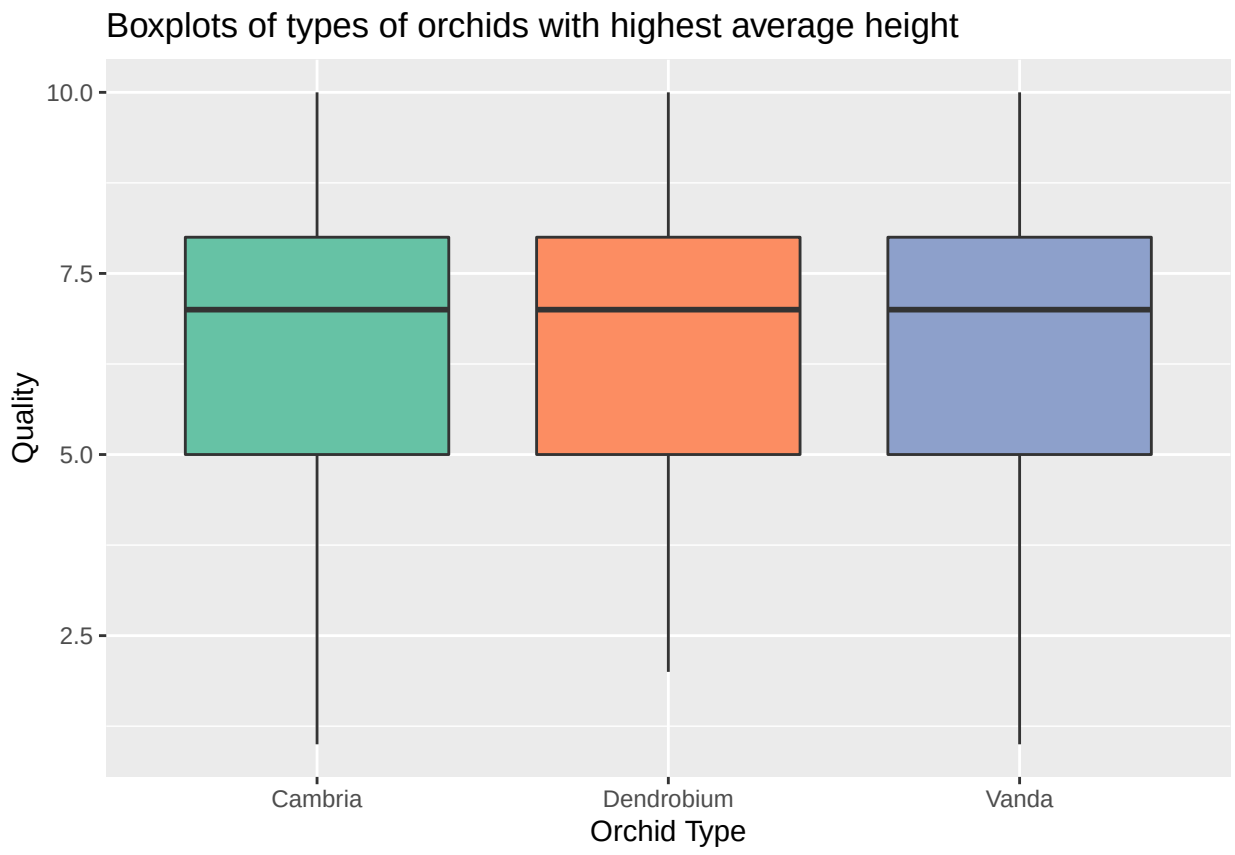
This plot suggests that temperature and height may have a weak positive correlation across each of the Orchid types plotted. We can have a look at the correlation coefficients to check this.

```
Orchids_Filtered %>%
  group_by(Type) %>%
  summarise("Correlation coefficient" = round(cor(Height, Temp),4))
```

```
## # A tibble: 3 x 2
##   Type       `Correlation coefficient`
##   <chr>                         <dbl>
## 1 Cambria                       0.181
## 2 Dendrobium                    0.124
## 3 Vanda                         0.206
```

The values are all roughly 0.1 - 0.2, so there looks to be a very weak positive correlation between temperature and height.

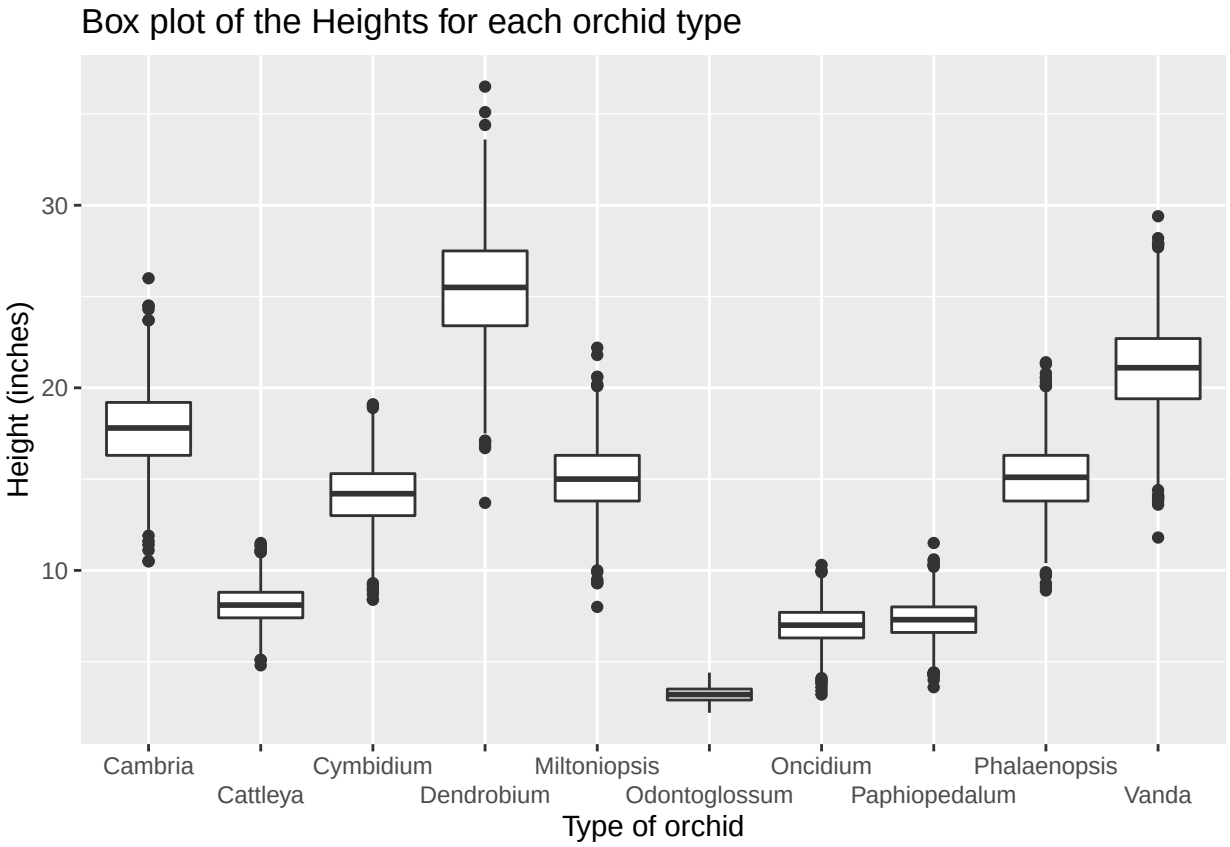
```
ggplot(Orchids_Filtered, aes(x = Type, y = Quality, fill = Type)) +
  geom_boxplot() +
  labs(title = "Boxplots of types of orchids with highest average height",
       x = "Orchid Type") +
  scale_fill_brewer(palette = "Set2") +
  theme(legend.position = "none")
```



The distribution of quality across these 3 types of orchids looks pretty much identical - this means the proportion of orchids that the shop buys will be roughly the same for each of these 3 types.

c) *Investigate differences between the types of orchids in terms of their distribution of height. Are there any differences in growing conditions?* [**5 marks**]

```
ggplot(Orchids, aes(x = Type, y = Height)) +
  geom_boxplot() +
  labs(title = "Box plot of the Heights for each orchid type",
       x = "Type of orchid", y = "Height (inches)") +
  scale_x_discrete(guide = guide_axis(n.dodge = 2))
```



This plot illustrates how the distributons of height vary across the types of plants. We can see that Dendrobium has the highest average height, with Odontoglossum with the lowest. There are types of plants with wider distributions (e.g. Dendrobium, Vanda) and others with a very narrow range of values (e.g. Ondontoglossum, Oncidium). Let's have a look at the growing conditions of the types of plants.

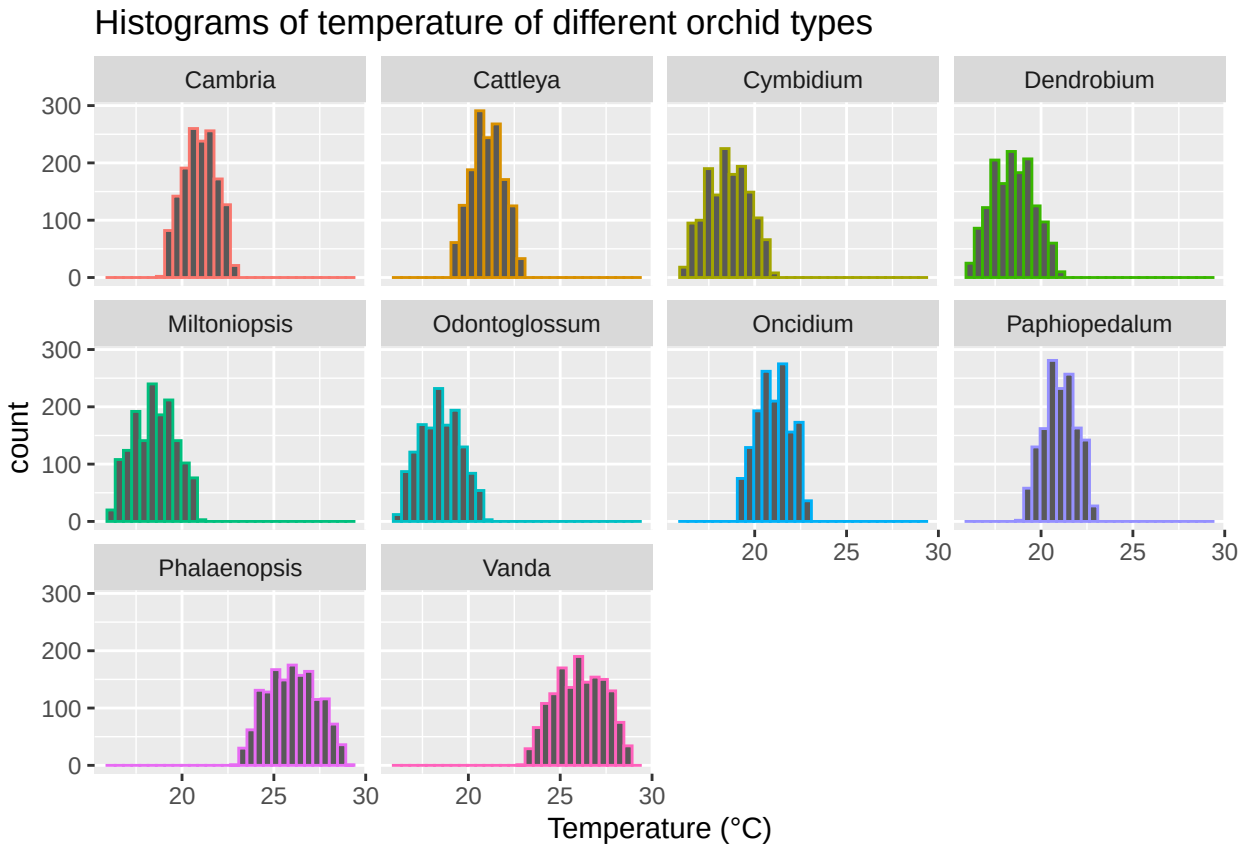
```
Orchids %>%
  select(Type,Phos, Potas, Sulf, Planting, Temp) %>%
  group_by(Type) %>%
  na_if(0) %>%
  summarise_all(list(avg = mean), na.rm = TRUE)
```

```
## # A tibble: 10 x 6
##    Type          Phos_avg Potas_avg Sulf_avg Planting_avg Temp_avg
```

```
##    <chr>            <dbl>     <dbl>    <dbl> <date>          <dbl>
##  1 Cambria           79.8      281.     36.1 2022-03-24       21.0
##  2 Cattleya          79.9      280.     36.3 2022-03-24       21.0
##  3 Cymbidium         80.2      280.     36.2 2022-03-25       18.5
##  4 Dendrobium        79.9      280.     36.3 2022-03-24       18.5
##  5 Miltoniopsis      80.3      279.     36.3 2022-03-25       18.5
##  6 Odontoglossum     80.1      280.     36.4 2022-03-24       18.5
##  7 Oncidium          80.2      280.     36.3 2022-03-25       21.0
##  8 Paphiopedalum     80.1      281.     36.2 2022-03-24       21.0
##  9 Phalaenopsis      79.4      280.     36.2 2022-03-24       26.0
## 10 Vanda             79.7      281.     36.3 2022-03-24       26.1
```

This table of averages suggests that temperature may be a growing condition to look into (as averages of the other variables are roughly the same). Looking at histograms of the other variables (not shown) reveals very similar shaped distributions with similar averages for all the types of orchids, so there doesn't seem to be any difference in these conditons.

```
ggplot(Orchids, aes(x = Temp, colour = Type)) +
  geom_histogram(show.legend = FALSE) +
  labs(title = "Histograms of temperature of different orchid types",
       x = "Temperature (°C)") +
  scale_fill_brewer(palette = "Set2") +
  facet_wrap(~Type)
```



We can see there are differences in the temperatures that the different types of orchids are grown in - the distributions all look roughly normally distributed with a range of means and variances. Note the orchid

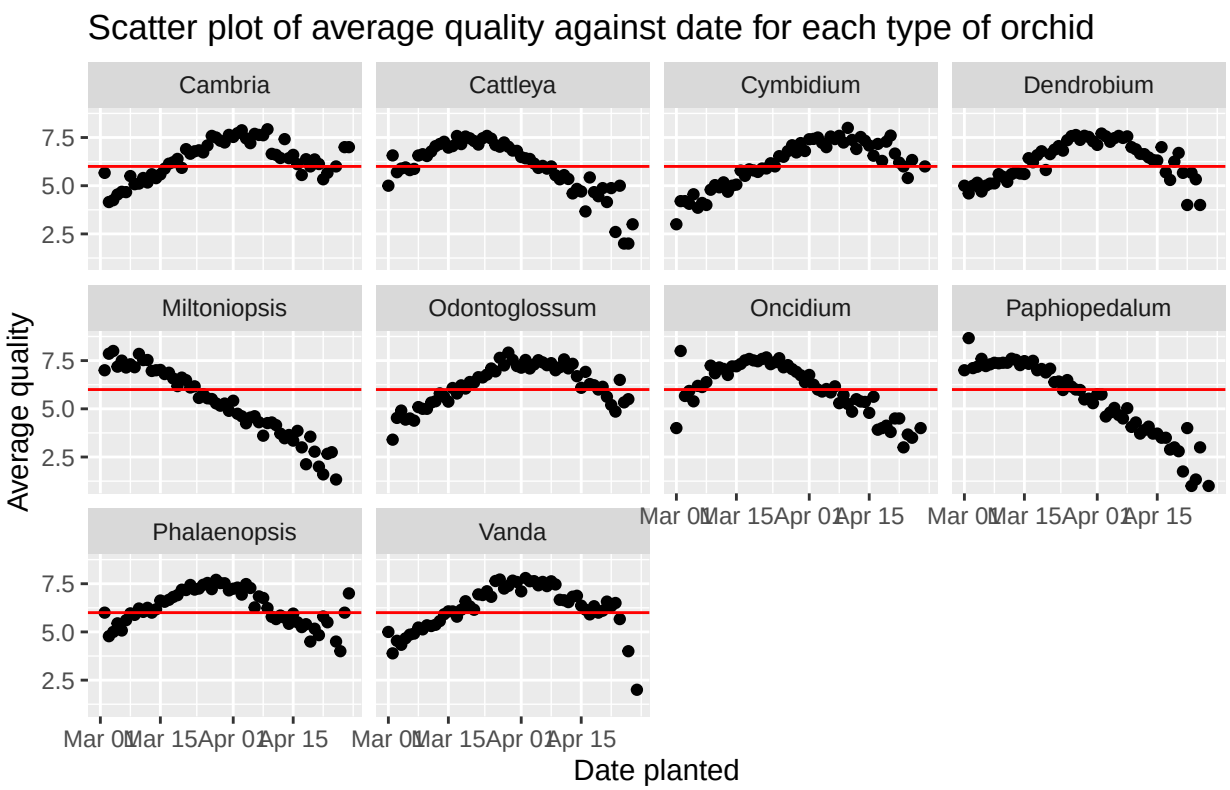
with the lowest quality distribution also had a lower temperature distribution, in fact comparing with the plot above, for the most part, the types of orchids that have higher temperature distributions also have higher quality distributions.

d) *The orchid grower wants to optimize the times at which the different types of orchids are transferred to individual pots. The aim is to have a large proportion of orchids being bought by the distributor. Use the data to advise the orchid grower on which two types of orchids they should plant first in 2023. When should the first orchid be planted? Discuss which assumption you make when basing your suggestions on the data.* [**5 marks**]

Calculate the average quality of each type of orchid at each date. Then analyse this metric to determine which of the plants should be planted first and at which date.

```
ByDateAverage <- Orchids %>%
  group_by(Type, Planting) %>%
  summarise("AvgQuality" = mean(Quality))
```

```
ggplot(ByDateAverage,aes(x = Planting, y = AvgQuality)) +
  geom_point() +
  geom_hline(yintercept = 6, colour = "red") +
  facet_wrap(~Type) +
  labs(title =
       "Scatter plot of average quality against date for each type of orchid",
       x = "Date planted", y = "Average quality",
       caption = "Note: points above the red line would be bought, while points
       on or below would not.")
```



Looking at this plot, we can see that all the plants have different time windows where their average quality is good enough to be bought by the distributor - (we want to chose the date corresponding to the maximum of these curves). We can also see that the windows for Miltoniopsis and Paphiopedalum are right at the start of the data, so these two orchids should be planted first. We can see that the peak of the Miltoniopsis' average quality looks to be at the start of the data range, while Paphiopedalum's follows shortly after. Therefore it is recommended that the orchid grower plant the Miltoniopsis first around 1st March 2023, with Paphiopedlum following shortly after. This assumes that the planting conditions in 2023 will be similar to those in 2022, so we would expect a similar distribution of quality.

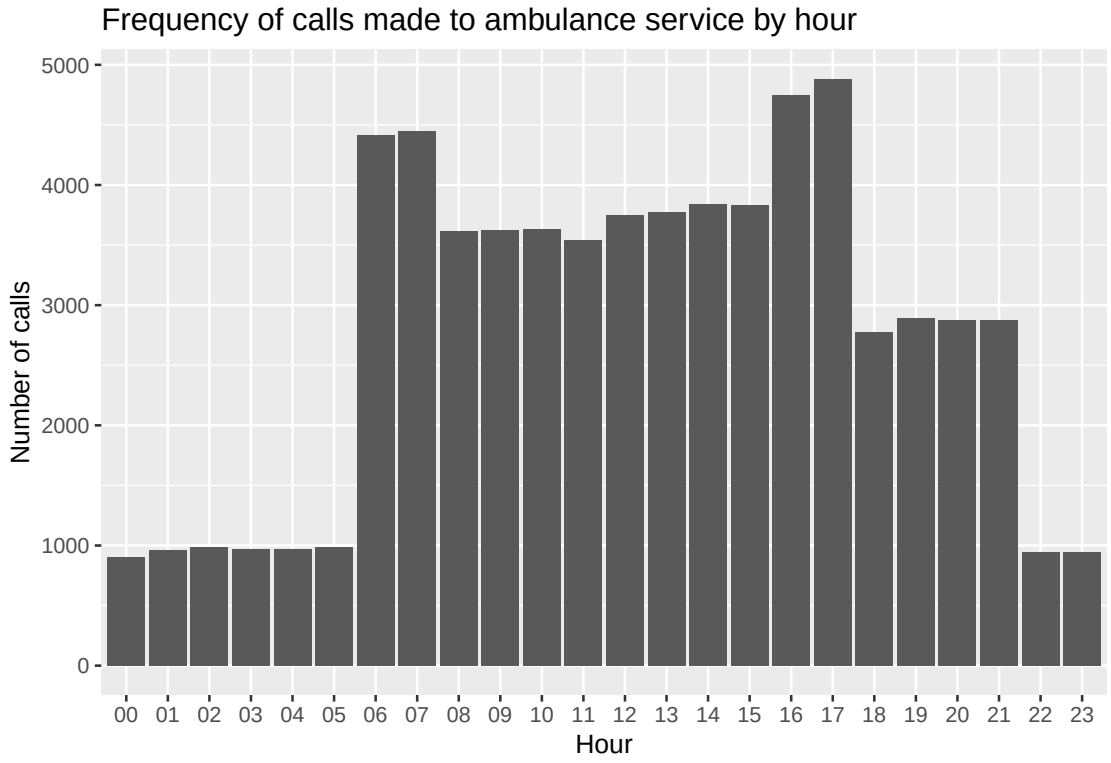
Question 2 [27 marks]

The country *Utopia* has collected data on their ambulance service and the patients admitted to the country's hospitals. The health department of Utopia has given you access to their data in the files "Ambulance.csv" and "Hospital.csv", and a data description is provided in the file "Data Descriptions.pdf". You are asked to consider the following tasks which are aimed towards analyzing the performance of their ambulance service and the factors influencing health outcomes:

a) *At which time of the day do we tend to see the highest frequency of calls to the ambulance service? Which proportion of calls leads to the patient being delivered to hospital?* **[4 marks]**

```
Ambulance <- read.csv("Ambulance.csv")
Ambulance$Call <- ymd_hms(Ambulance$Call)
Ambulance$Arrival <- ymd_hms(Ambulance$Arrival)
Ambulance$Hospital <- ymd_hms(Ambulance$Hospital)
```

```
ggplot(Ambulance, aes(x = format(as.POSIXct(Call), format = "%H"))) +
  geom_bar() +
  labs(title = "Frequency of calls made to ambulance service by hour",
       x = "Hour", y = "Number of calls")
```



Grouping the data by hour, the largest volume of calls is between 16:00 and 18:00.

```
1 - (sum(is.na(Ambulance$Hospital)) / nrow(Ambulance))
```

```
## [1] 0.8002888
```

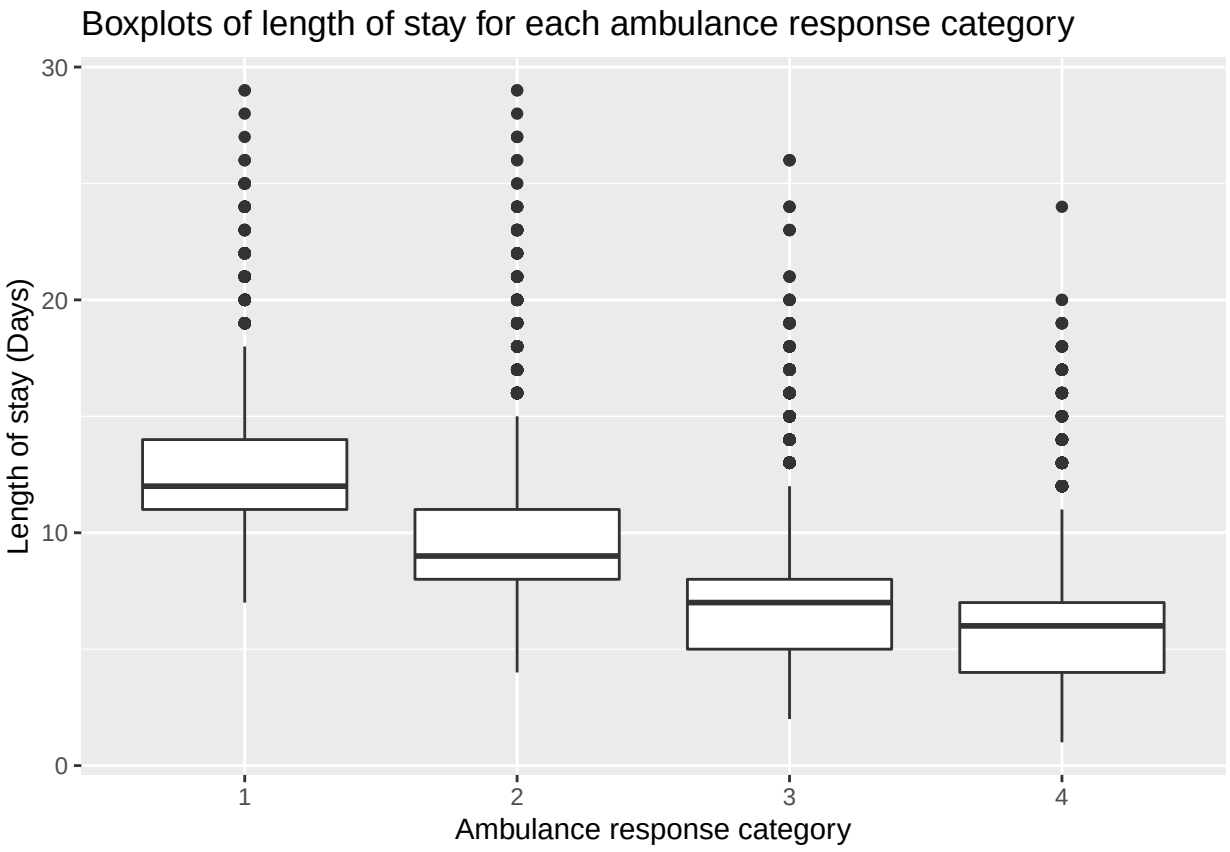
We see that around 80% of calls lead to a patient going to hospital.

b) *How does the length of stay in hospital and the probability of discharge from hospital vary across the four ambulance response categories? Here, ambulance response category refers to that at the time of arrival of the ambulance.* [**4 marks**]

Join the ambulance and hospital data frames along PatientID, include only patients who made a call to the ambulance and then went to hospital.

```
Hospital <- read.csv("Hospital.csv")
AmbHos <- left_join(filter(Ambulance, !is.na(Hospital)), Hospital,
                    by = "PatientID")
```

```
ggplot(AmbHos, aes(x = as.factor(Category2) , y = Length)) +
  geom_boxplot() +
  labs(title =
       "Boxplots of length of stay for each ambulance response category",
       x = "Ambulance response category", y = "Length of stay (Days)")
```




```
AmbHos %>%
  group_by(Category2) %>%
  summarise("Average length of stay" = round(mean(Length), 2) ,
            "Probability of discharge" = round(1 - mean(Outcome),4))
```

```
## # A tibble: 4 x 3
##   Category2 `Average length of stay` `Probability of discharge`
##       <int>                    <dbl>                       <dbl>
## 1         1                    13.0                        0.850
## 2         2                     9.62                       0.921
## 3         3                     7.14                       0.972
## 4         4                     6.12                       0.990
```

We can see that the average length of stay is shorter and the probability of discharge is greater for less urgent response categories. A box plot of length of stay shows that the length of stay for each category has similar looking distributions, just shifted by a value.

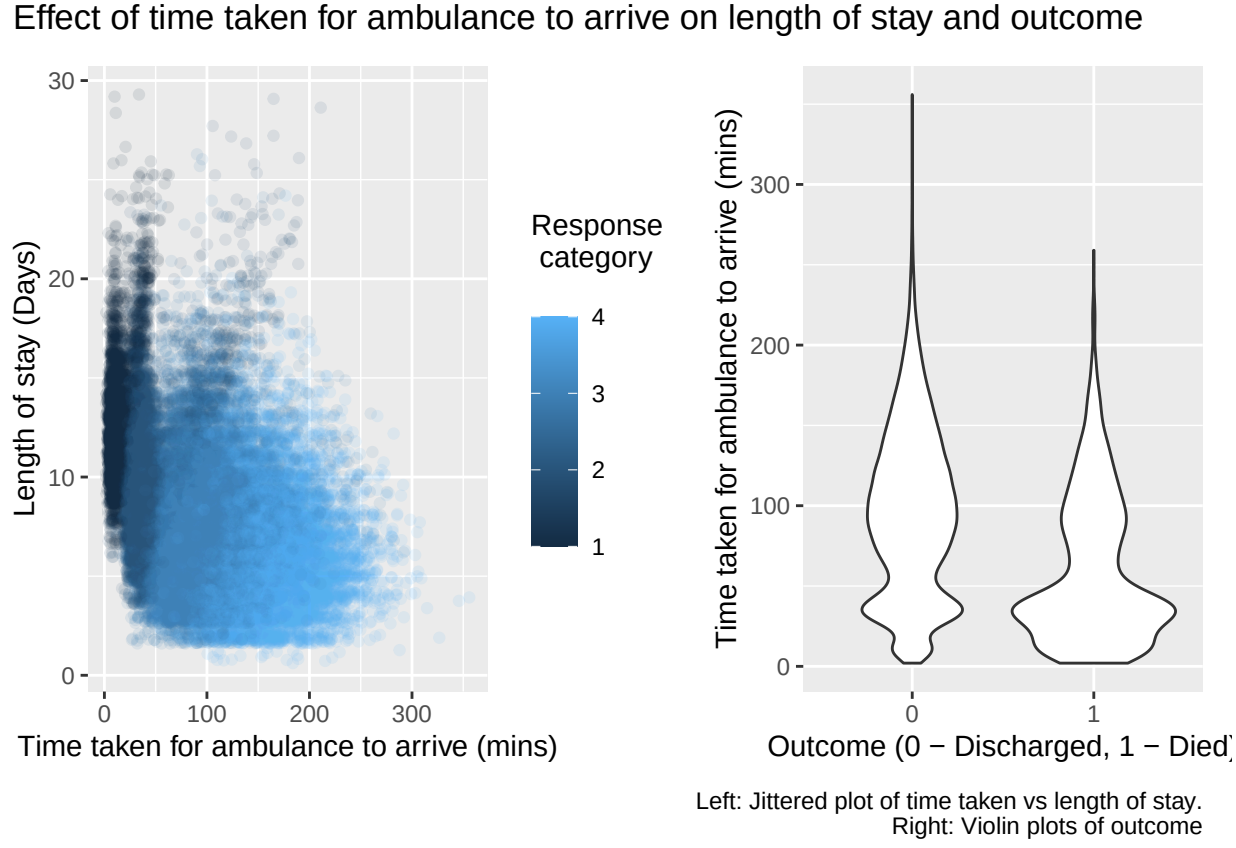
c) *Does the data suggest that the length of stay in hospital and the risk of death increase with the time until the ambulance arrives, i.e, the length of time between calling the ambulance service and the ambulance arriving?* [**5 marks**]

```
AmbHos_TDiff <- AmbHos %>%
  mutate("TimeForAmb" = Arrival - Call, .keep = "all")
```

```
p1 <- ggplot(AmbHos_TDiff, aes(x = TimeForAmb, y = Length,
                               colour = Category2)) +
  geom_jitter(alpha = 0.1) +
  labs(x = "Time taken for ambulance to arrive (mins)",
       y = "Length of stay (Days)",
       colour = "Response \n category \n")


p2 <- ggplot(AmbHos_TDiff, aes(x = as.factor(Outcome), y = TimeForAmb)) +
  geom_violin() +
  labs(x = "Outcome (0 - Discharged, 1 - Died)",
       y = "Time taken for ambulance to arrive (mins)")


p1 + p2 + plot_annotation(title =
"Effect of time taken for ambulance to arrive on length of stay and outcome",
                  caption = "Left: Jittered plot of time taken vs length of stay.
                  Right: Violin plots of outcome")
```



We can see from the left plot that there seems to be some negative impact of a longer time taken for the ambulance to come on length of stay. From the right plot, the risk of death is much larger at lower times taken - this can be explained by more life-threatening calls taking priority (these calls would likely lead to longer stays with higher risk of death).

d) *Make up your own question and answer it. Your question should be aimed towards understanding the factors influencing length of stay in hospital / health outcome. Originality will be rewarded.* [**7 marks**]

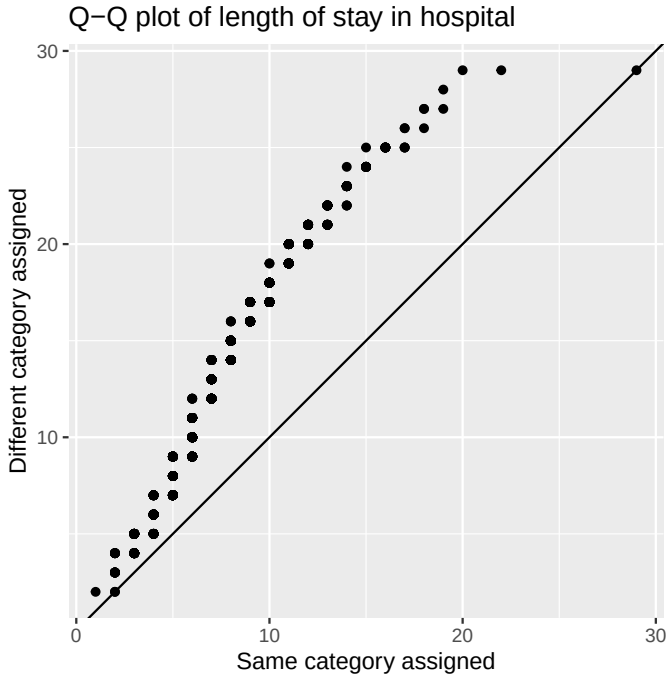
When a patients condition worsened between the time of the call and the ambulance arriving (corresponding to Category1 > Category2), is there a difference in the distributions of length of stay in hospital and health outcome compared to the rest of the data?

```
AmbHos_CatDiff <- AmbHos %>%
  filter(Category1 > Category2) %>%
  select(PatientID, Category1, Category2, Length, Outcome)

AmbHos_NoCatDiff <- AmbHos %>%
  anti_join(AmbHos_CatDiff, by = "PatientID") %>%
  select(PatientID, Category1, Category2, Length, Outcome)

qq.out <- qqplot(x = AmbHos_NoCatDiff$Length,
                 y = AmbHos_CatDiff$Length, plot.it = FALSE)
qq.out <- as.data.frame(qq.out)

ggplot(qq.out, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1) +
  coord_fixed(ratio = 1) +
  labs(title = "Q-Q plot of length of stay in hospital",
       x = "Same category assigned",
       y = "Different category assigned")
```



We can see that as we move to larger number of days in hospital, the distributions stray away from each other - the distribution of length of stay in hospital when a different category was assigned looks to have much larger values in the middle quantiles and upper quantiles - therefore the distributions may be different. The patient is more likely to stay a longer duration if their condition worsened whilst waiting for the ambulance.

```
round(1- mean(AmbHos_NoCatDiff$Outcome),4)
```

```
## [1] 0.9619
```

```
round(1- mean(AmbHos_CatDiff$Outcome),4)
```

```
## [1] 0.8597
```

We see that the probability of being discharged when the same category was assigned when the ambulance arrived (96.19%) is quite a bit higher than when the patients condition worsened before the ambulance's arrival (85.97%).

e) *Write a short (two paragraphs) report about the findings of your analysis in parts a-d. The report should be readable for people without data science knowledge. Make it sound interesting and state possible recommendations that may be of interest to Utopia's health department.* [**7 marks**]

From looking at the data supplied by Utopia's health department, we were able to identify different trends in the ambulance service. We found that most calls for the ambulance tend to be made in the late afternoon (4-6pm) and that 4 in 5 calls lead to the patient being taken to hospital. Utopia's ambulance response times were faster for patients with more urgent conditions, these patients tended to spend longer in the hospital and were at a larger risk of death. Interestingly, patients whose condition had worsened in between the time they made the call and the time the ambulance arrived have a much larger stay in hospital compared with those whose condition remained the same, they also had a higher risk of death.

Utopia's health department could do more exploration of the data, especially on the relationships between factors such as Age, BMI and whether or not the patient had an operation. It could be interesting to try to attempt to set a target response time given the medical information that the patient can provide when calling an ambulance.