MA20277 Coursework 1

Ariana Vass 23111

```
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]**

I will load the data and it by changing the format of Planting to a Date:

```
Orchids <- read.csv("Orchids.csv")
Orchids$Planting <- as_date(Orchids$Planting, format="%Y-%m-%d")
```

This extracts the first two rows of the data set:

```
slice_head(Orchids,n=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
```

Extracting the minimum and maximum levels of Phosphate, Potassium, and Sulphur: (I have chosen to print my summary and units to avoid repeating the data)

```
cat(" Observed phosphate levels:\n ",
    "min: ", min(Orchids$Phos [Orchids$Phos> 0]), "ppm, max: ",
    max(Orchids$Phos), "ppm")
```

```
##  Observed phosphate levels:
##   min:  46 ppm, max:  130 ppm
```

```
cat(" Observed potassium levels:\n ",
    "min: ", min(Orchids$Potas [Orchids$Potas > 0]), "ppm, max: ",
    max(Orchids$Potas),"ppm")
```

```
##  Observed potassium levels:
##   min:  195 ppm, max:  385 ppm
```

```
cat(" Observed sulphur levels:\n ",
    "min: ", min(Orchids$Sulf [Orchids$Sulf > 0]), "ppm, max: ",
    max(Orchids$Sulf),"ppm")
```

```
##  Observed sulphur levels:
##   min:  28 ppm, max:  46 ppm
```

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**]

First we need to find the 3 types of orchid with the highest average height, we do this by grouping by the variable Type and calculating the average height of them:

```
Orchids %>%
  group_by(Type) %>%
  summarise(average_height = mean(Height)) %>%
  arrange(desc(average_height)) %>%
  slice_head(n=3)
```

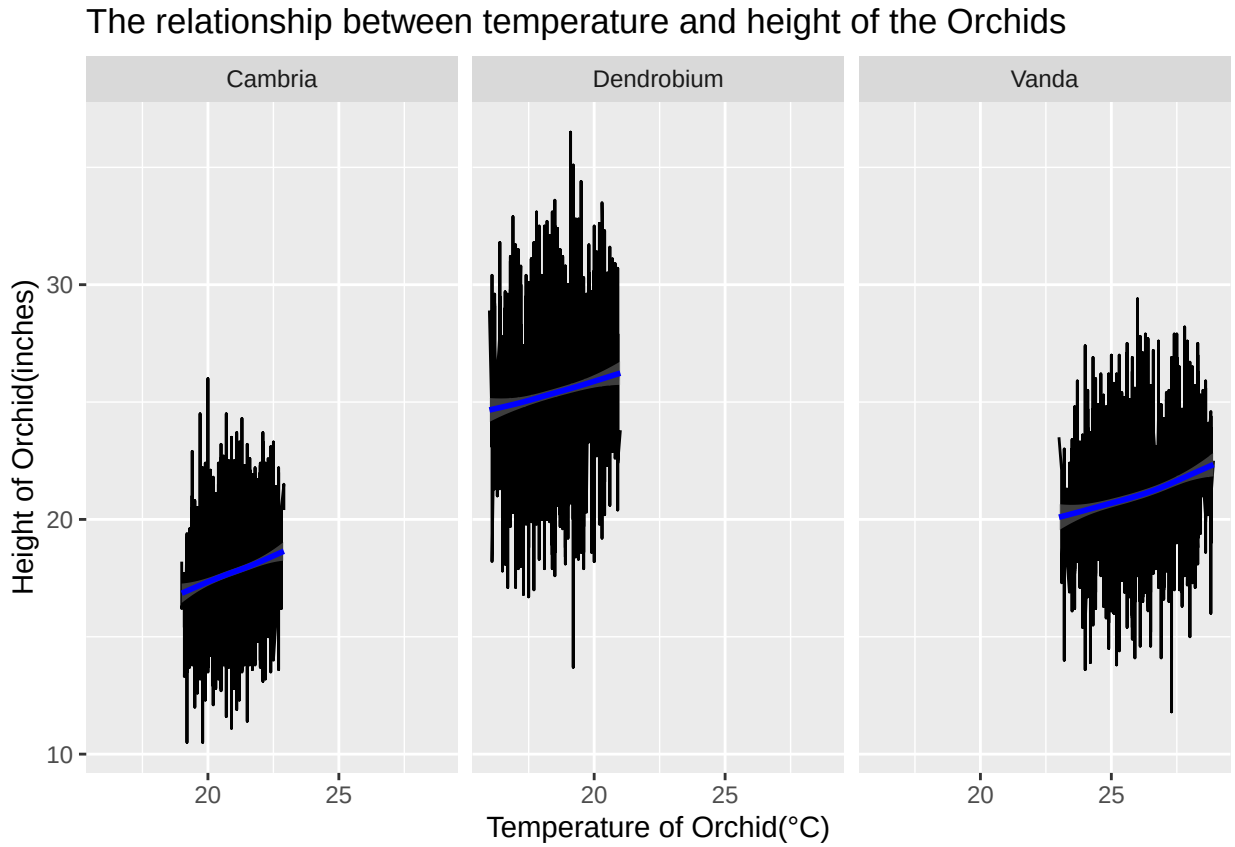
```
## # A tibble: 3 x 2
##   Type       average_height
##   <chr>               <dbl>
## 1 Dendrobium           25.4
## 2 Vanda                21.1
## 3 Cambria              17.8
```

We find that the three types with the highest average height are Dendrobium, Vanda, and Cambria.

To explore the relationship of temperature and plant height in each of these, we can make 3 line plots:

```
Tallest_Orchids <- filter(Orchids,
                          Type == "Dendrobium" | Type == "Vanda" | Type == "Cambria")
ggplot(Tallest_Orchids, aes(x = Temp,y = Height)) +
  geom_line() + facet_wrap(~Type) + geom_smooth( color="blue" ) +
  ggtitle("The relationship between temperature and height of the Orchids") +
  labs(x = "Temperature of Orchid(\u00B0C)", y = "Height of Orchid(inches)")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

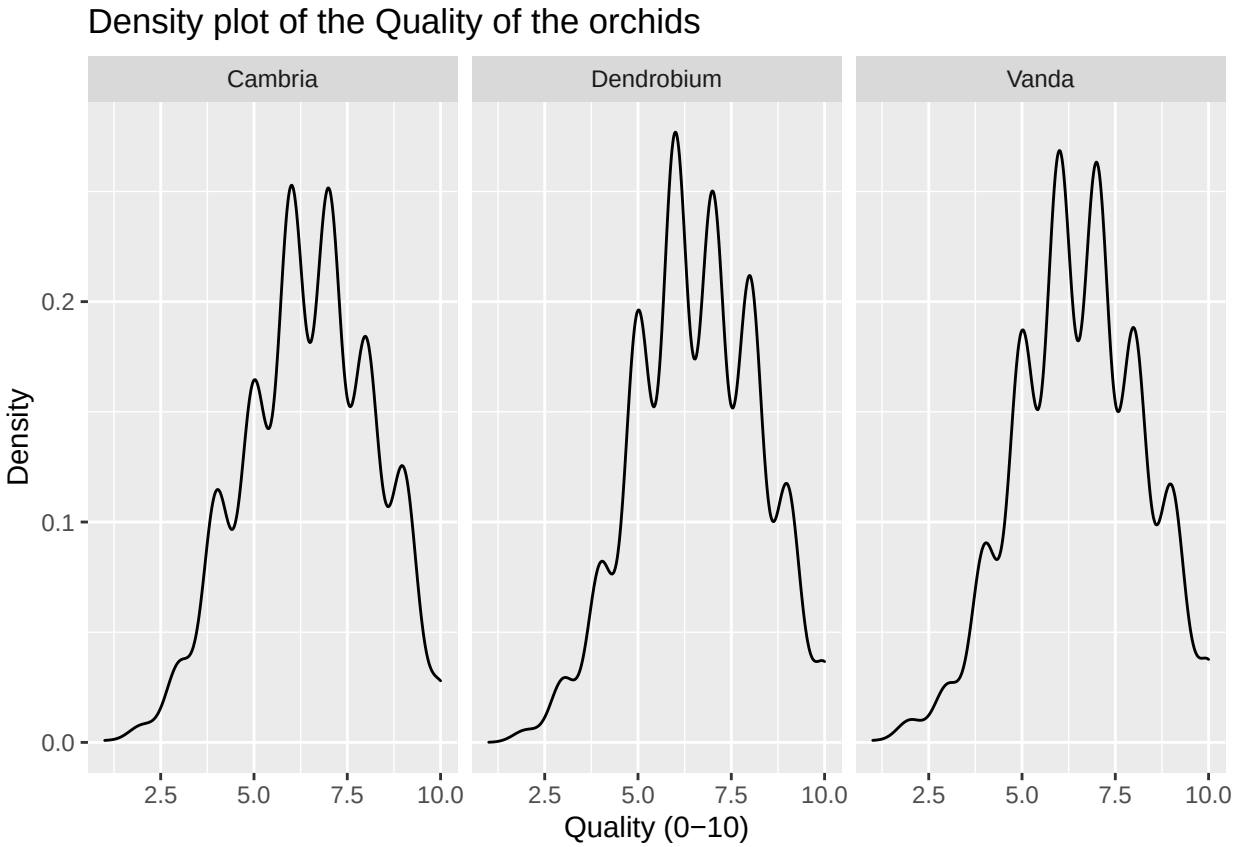


I added "geom_smooth()" to make it easier to interpret the overall correlation between the two variables. In general, for each of these graphs, we find a positive correlation between the temperature and the height of the Orchids. We also see that they are each grown in different kinds of temperatures; Cambria are grown between roughly 19 - 23°C, Dendrobium are grown between roughly 6 - 21°C, and Vanda are grown between roughly 23 - 29°C

We do not have more data for different temperature intervals. Therefore, we are unable to see how temperatures outside of the data set affect the height of these varieties of orchid. For example we cant see how too hot conditions affect the height.

To investigate how these 3 varieties compare regarding their quality, I am going to use a density plot of their scores given by the distributor and facets for each of the 3 varieties:

```
ggplot(Tallest_Orchids,aes(x= Quality)) +
  geom_density() + facet_wrap(~Type) +
  ggtitle("Density plot of the Quality of the orchids") +
  labs(x="Quality (0-10)", y = "Density")
```

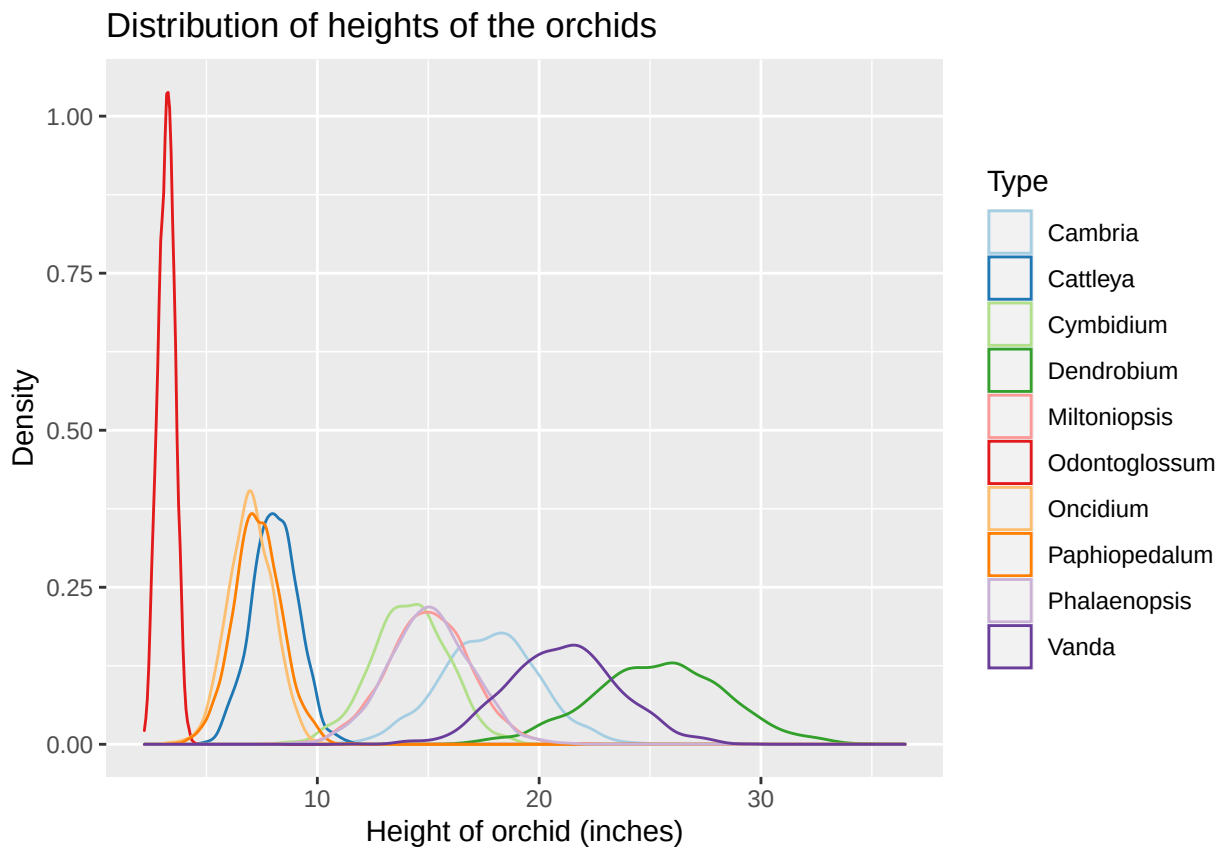


For all 3 varieties, we see high density spikes at 6 and then at 7. We can see that the Dendrobium variety has the highest density of orchids scoring higher than 6 and so, out of the 3, will have the largest proportion of orchids bought by the distributor, i.e. they are the best quality of the 3. The Vanda variety will have had the lowest proportion of orchids bought with the highest density of orchids scoring 6 or below, meaning they are the worst quality of the 3.

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

To investigate the distribution of height, I am going to use a density plot with each type of orchid being a different line colour, in order to make comparison easier:

```
ggplot(Orchids, aes(x = Height)) +
  geom_density(aes(color = Type)) +
  ggtitle("Distribution of heights of the orchids") +
  labs(x= "Height of orchid (inches)",y = "Density") +
  scale_color_brewer( palette="Paired" )
```

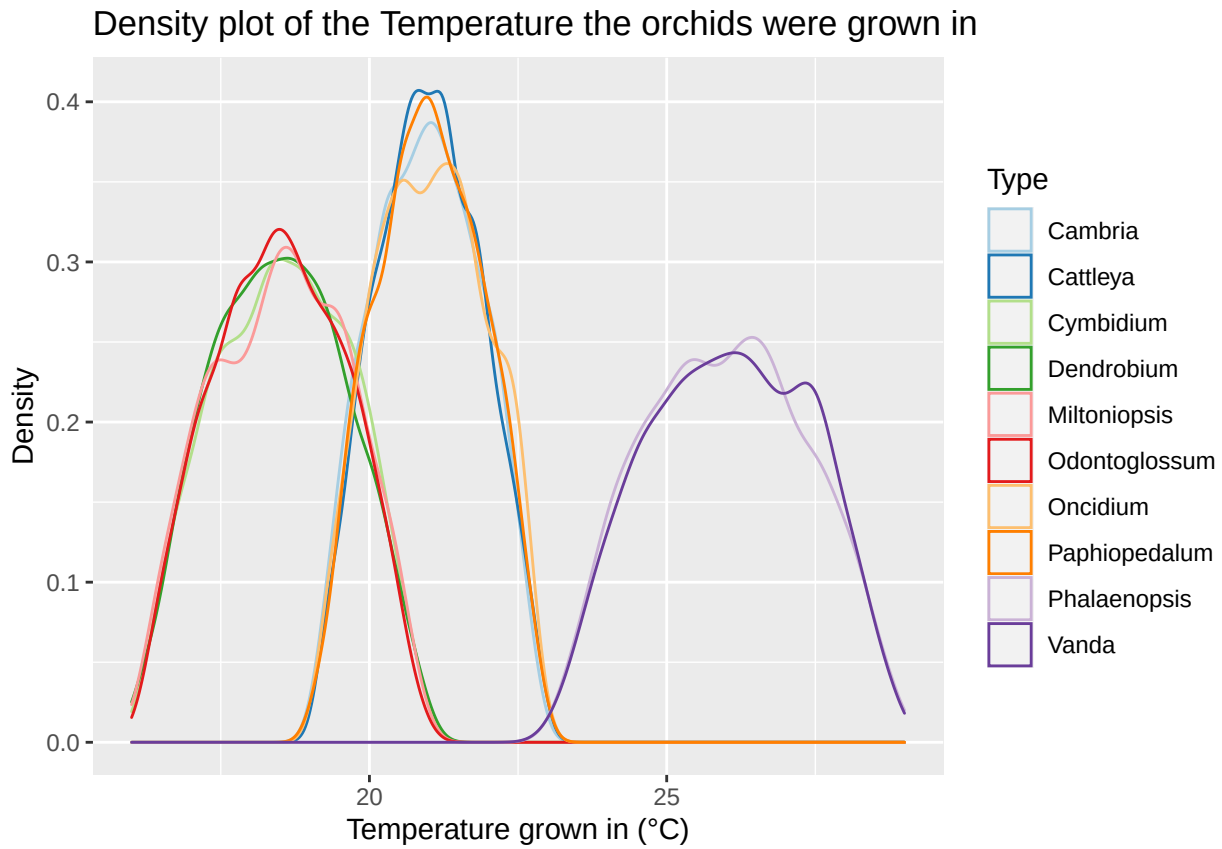


I decided to change the colour palette to more easily tell the difference between the orchids.

- Here we see that the variety Odontoglossum grow a smaller height on average than the other varieties, most of them growing to only around 3-4 inches, they also have the smallest range in height.
- Dendrobium grow to the tallest height of all the varieties, at an aproximate average of 25 inches, and has the largest range in height.
- The varieties Oncidium, Paphiopedalum, and Cattleya have very similar distributions of height, growing to an average of around 6-8 inches tall.
- The varieties Cymbidium, Miltoniopsis, and Phalaenopsis also have very similar distributions of height, growing to an average of around 13-15 inches.

To investigate the differences in their growing conditions, I am going to create a plot looking at the temperature that each variety is grown in:

```
ggplot(Orchids, aes(x = Temp)) +
  geom_density(aes(color = Type)) +
  ggtitle("Density plot of the Temperature the orchids were grown in") +
  labs(x= "Temperature grown in (\u00B0C)",y = "Density") +
  scale_color_brewer( palette="Paired" )
```

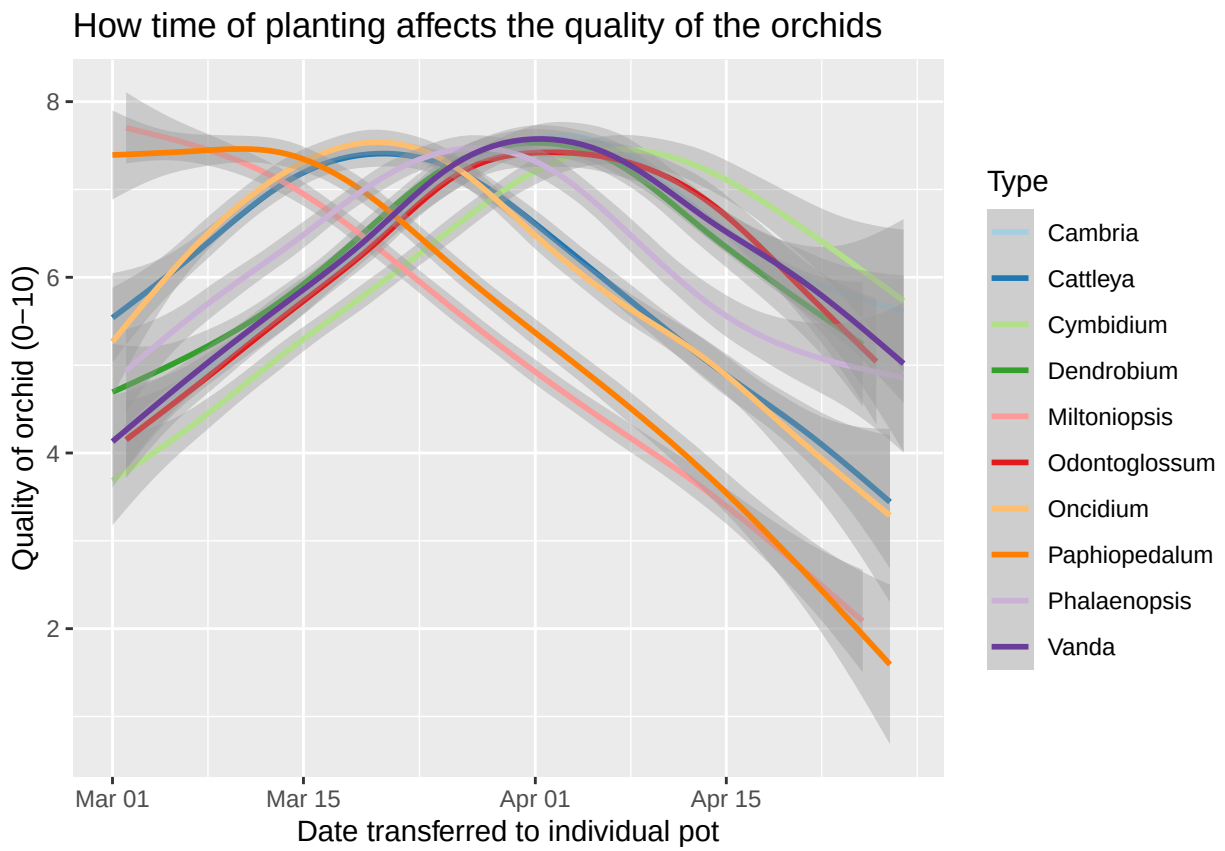


* We see that there are 3 ranges of temperatures that the orchids are grown in: the highest density of orchids being grown at around 21 degrees. Odontoglossum, which are the shortest, are grown in the lowest range of temperatures(average of ~17 degrees). The Vanda variety is the second tallest, and is grown in the highest temperature range(average of ~26degrees). Otherwise there is not enough similarity between height and temperature for correlation to be found. 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**]

To Determine the best times are to plant each variety of orchid, I am going to look at the variables Planting and Quality, to see how the date it's transferred to an individual pot affects the quality of the orchid, because we know that the distributor will buy all the orchids with a quality score above 6.
I will use a scatter graph to investigate the correlation between Quality and Planting. However, I have chosen to hide the points and use smooth lines with different colours for each variety of orchid:

```
ggplot(Orchids, aes(x = Planting, y = Quality)) +
  geom_smooth(aes(color = Type)) +
  ggtitle("How time of planting affects the quality of the orchids") +
  labs(x= "Date transferred to individual pot", y = "Quality of orchid (0-10)") +
  scale_color_brewer(palette = "Paired")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



I chose not to display the points because it made reading the graph confusing. I used the same colour palette as in the previous graph for consistency and to make the graph more legible. With the scatter graph and lines to show correlation, we can immediately see that there are two varieties that have peak quality scores when potted in early March: Miltoniopsis and Paphiopedalum. Therefore, I would recommend that Miltoniopsis and Paphiopedalum should be the first two orchids to be planted in 2023. This would increase the proportion of orchids bought by the distributor.

I would suggest that the first orchid to be planted should be the Miltoniopsis variety and should be planted on the first of March. However, we don't have data from before the first of March and so cannot see if this variety peaked before this point. The other varieties peaked with scores of 7.5 and the Miltoniopsis variety gets a score of above 7.5 when planted on the 1st of March so I have made the assumption that this is the peak time to plant as we are missing data before this point.
I have also made the assumption that the weather and other conditions will be the same in Spring 2023 as they were in Spring 2022, as weather conditions could heavily affect the quality of the orchids if the conditions are different. With the data we have, my best recommendation is for the first orchid to be planted on the first of march to increase the proportion of orchids bought by the distributor.

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:

First, I am going to load the data and change the data types of Call, Arrival, and Hospital. This is in the Tidy data format:

```
Ambulance <- read.csv("Ambulance.csv")
Hospital <- read.csv("Hospital.csv")
Ambulance$Call <- as_datetime(Ambulance$Call, format = "%Y-%m-%d %H:%M:%S")
Ambulance$Arrival <- as_datetime(Ambulance$Arrival, format = "%Y-%m-%d %H:%M:%S")
Ambulance$Hospital <- as_datetime(Ambulance$Hospital, format = "%Y-%m-%d %H:%M:%S")
```

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**]

To find the time at which the most calls are made to the ambulance service I need to extract the time from Call and group by time and summarise how many calls are received at each time period. I have chosen to extract the hour that the ambulance service was notified in rather than precise time because I felt that extracting such specific data would not be helpful to the health department, whereas they could act more easily on hourly data.

```
Ambulance %>%
  mutate(Time_of_day = format(Call, format = "%H")) %>%
  group_by(Time_of_day) %>%
  summarise("Number of calls" = n()) %>%
  arrange(desc(Time_of_day)) %>%
  slice_head(n=1)
```

```
## # A tibble: 1 x 2
##   Time_of_day `Number of calls`
##   <chr>                   <int>
## 1 23                        947
```

The Time of day that we see the highest frequency of calls to the ambulance service is during the hour 23:00-23:59. From this I would suggest thinking ambulance crew on duty from 23:00
To find the proportion of calls that leads to the patient being delivered to hospital, I am going to find the proportion of NA values in the variable Hospital and the proportion of calls leads to the patient being delivered to hospital is 1-that :

```
1 - mean(is.na(Ambulance$Hospital))
```

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

We see that 80% of the calls lead to the patient being delivered 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**]

First we need to merge the two data sets, joining by PatientID. For this question, we only want the rows where there is data in both data sets so I am going to use "inner_join": (we only want data for calls that resulted in hospitalisations)

```
Medical_records <- inner_join(Ambulance,Hospital, by = c("PatientID" = "PatientID"))
```

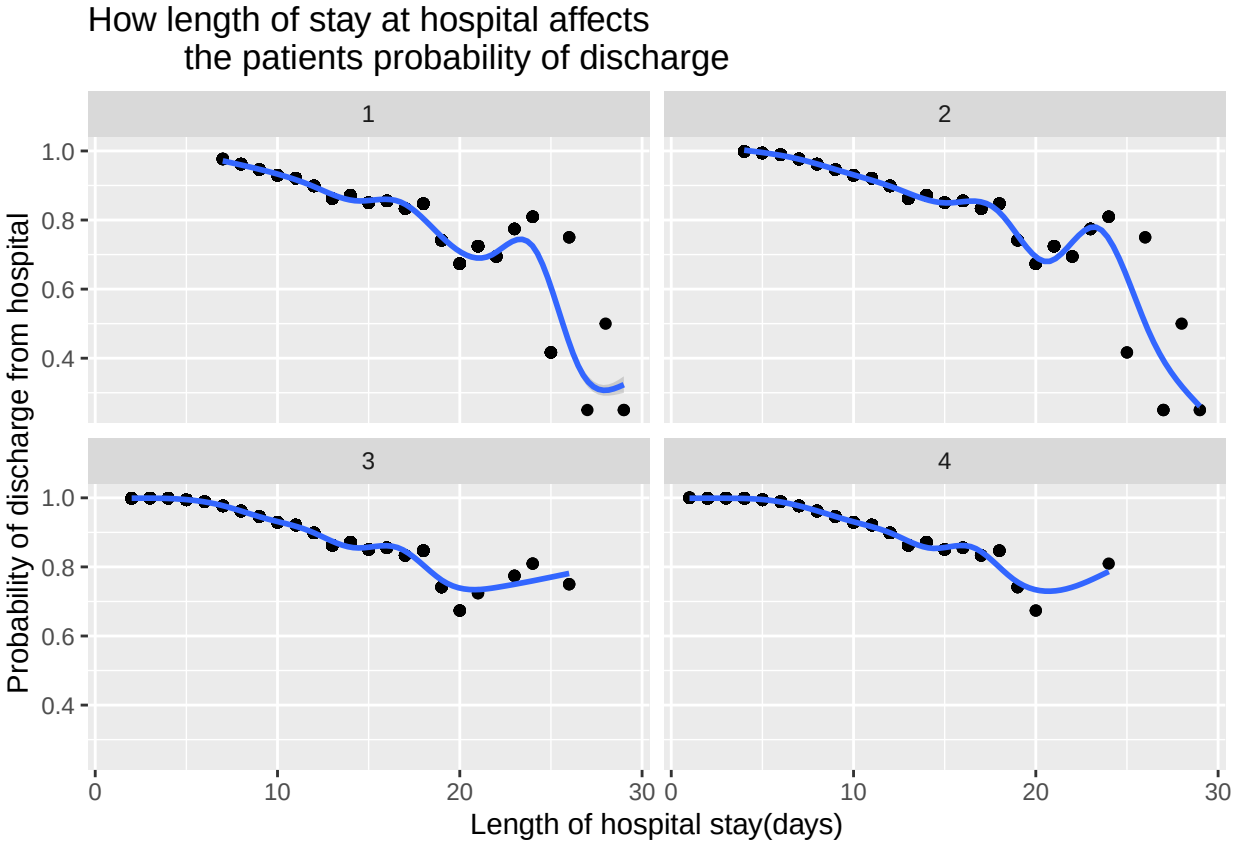
Now to investigate how the length of hospital stay and the probability of discharge vary across the four ambulance response categories. Firstly, I have chosen to group by length of stay and create a new variable to illustrate the probability of discharge for each length of stay:

```
Medical_records_by_length <- Medical_records %>%
  group_by(Length) %>%
  mutate(Probability_discharge = 1 - mean(Outcome)) %>%
  ungroup()
```

Now we want to plot the length of stay against the probability of discharge and separate into different facets for each of the four ambulance response categories:

```
ggplot(Medical_records_by_length,aes(x = Length, y = Probability_discharge)) +
  geom_point() + facet_wrap(~Category2) + geom_smooth() +
  ggtitle("How length of stay at hospital affects
          the patients probability of discharge") +
  labs(x= "Length of hospital stay(days)", y = "Probability of discharge from hospital")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Overall, we see that generally the longer the stays in hospital, the lower the probability of discharge. We also see, for all the categories, at around 20 days, there is an increase in the probability of discharge. Category 3&4 conditions rarely stay in the hospital for more than 25 days.
We also see that for the more serious categories, 1 & 2, the overall probability of discharge is lower. Furthermore, the longer the patient stays in hospital, the less likely that they are discharged.

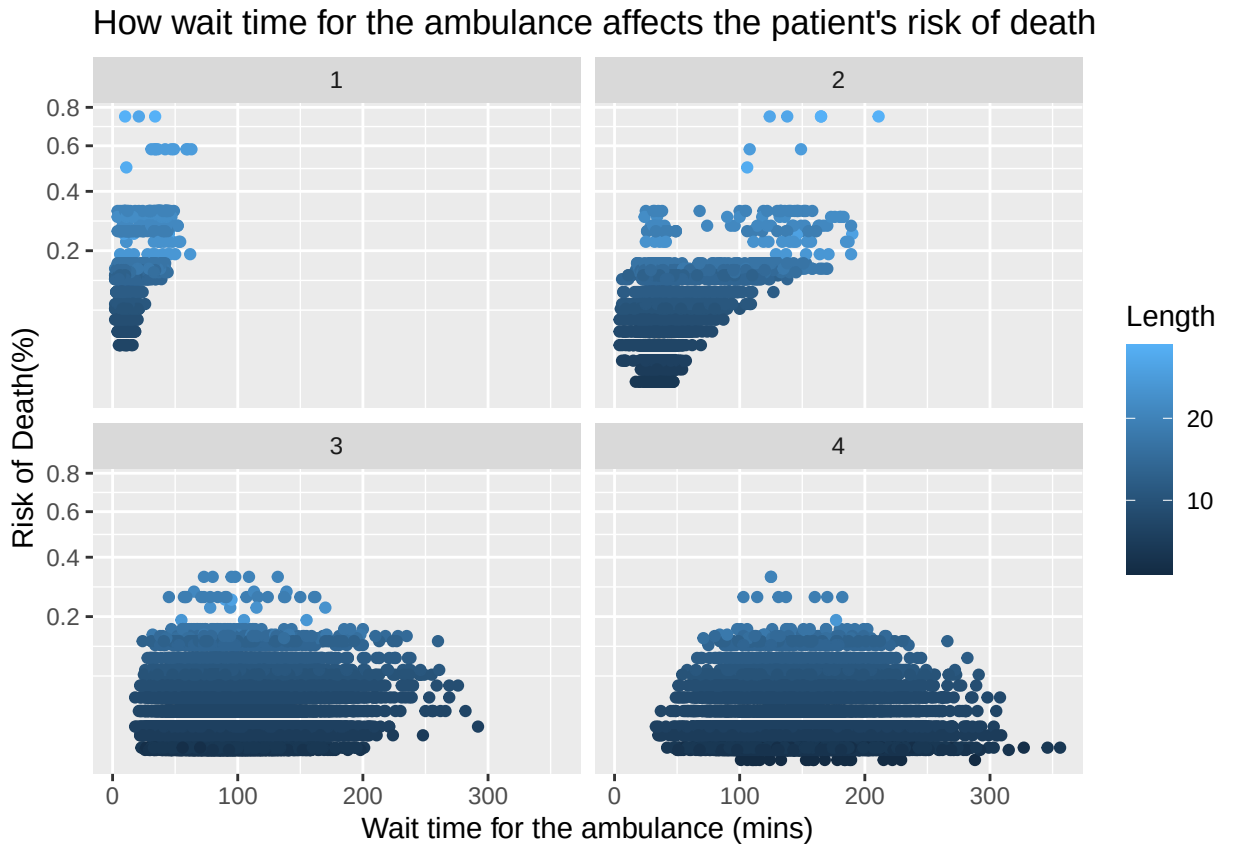
c) *Does the data suggest that the length of stay in hospital and the risk of death increase with the time until the ambulance arrives, i.e, the length of time between calling the ambulance service and the ambulance arriving?* **[5 marks]**

First we need to extract the time between calling the ambulance service and the ambulance arriving, then assign it to a new variable, Wait_time. I have decided to plot length of stay in hospital and risk of death together. The position of the points showing risk of death and the colour showing the length of stay in the hospital:

```
Medical_records_wait_time <- Medical_records %>%
  mutate(Wait_time = Arrival - Call) %>%
  group_by(Length) %>%
  mutate(Risk_of_death = mean(Outcome)) %>%
  ungroup()

ggplot(Medical_records_wait_time, aes(x = Wait_time, y = Risk_of_death))+
  geom_point(aes(color = Length)) + scale_y_sqrt() + facet_wrap(~Category2) +
  ggtitle("How wait time for the ambulance affects the patient's risk of death") +
  labs(x= "Wait time for the ambulance (mins)", y = "Risk of Death(%)")
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



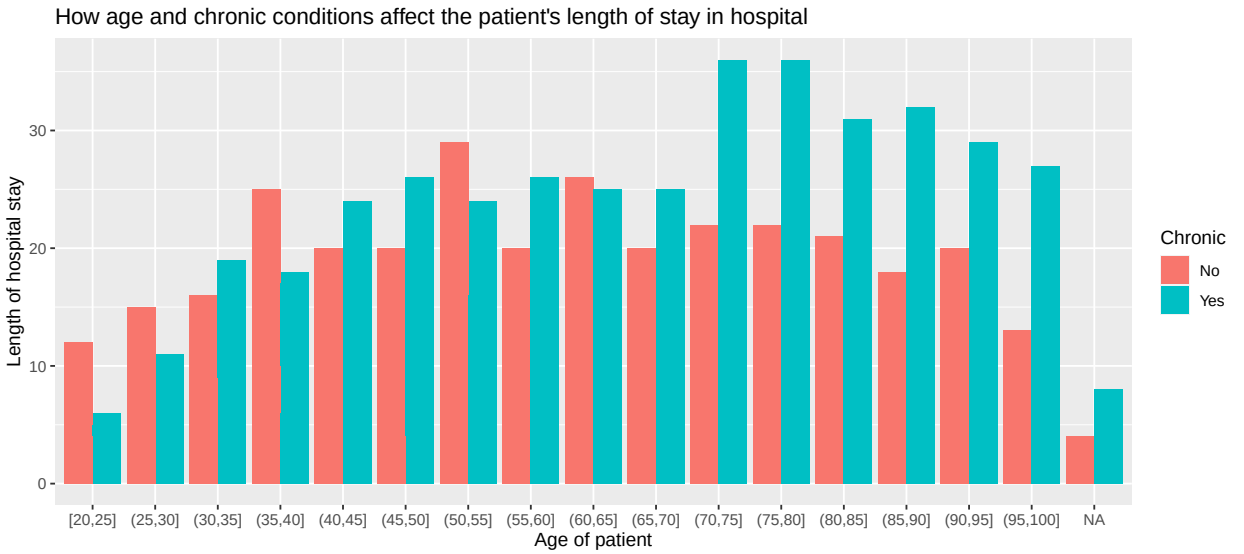
I chose to change the scale of the y-axis on the square-root scale, because it was hard to interpret when the points were all gathered together at 0 on the x-axis. I also chose to separate by category.
For both Length of stay and Risk of Death, in categories 3 & 4, we see a weak negative correlation between them and the time waiting for the ambulance. This means that the wait time for the ambulance doesn't affect the patient's recovery for less serious conditions. These patients are waiting a lot longer than the other two categories. For category 1 we see that these patients are waiting the shortest and a weak positive correlation between the risk of death and length of stay, and wait time for the ambulance. The Ambulance crews are prioritising these serious cases since risk of death is much higher.

We see for category 2, there is a stronger positive correlation between the risk of death and length of stay, and wait time for the ambulance. This shows that these patients are being negatively impacted by having to wait longer of the ambulance. Therefore, only for category 2 do we see that the length of stay in hospital and the risk of death increase with the time until the ambulance arrives.

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**]

I have decided to investigate how the age of patient and whether they have a chronic condition affect the length of stay. I predict that patients with chronic conditions will generally need to stay longer than ones that don't, as they need to be more closely monitored, and that the older the patient, the longer their stay. I'm going to look at just the data set Hospital because Medical_records doesn't contain the patients who didn't get to the hospital via ambulance.
To investigate this, I need to turn chronic into a characters instead of integers, for ease of understanding, I will change them to Yes and No:

```
Hospital$Chronic <- case_when(
  Hospital$Chronic == 1 ~ "Yes",
  Hospital$Chronic ==0 ~ "No"
)
Segment = factor(cut(Hospital$Age,breaks = seq(20,100,5),include.lowest = TRUE))
ggplot(Hospital, aes(x = Segment, y = Length, fill = Chronic)) +
  geom_col(position = "dodge") +
  ggtitle("How age and chronic conditions affect the patient's length of stay in hospital") +
  labs(x = "Age of patient", y = "Length of hospital stay")
```



I decided to use a bar chart to illustrate how age and chronic conditions affect the length of hospital stay. I segmented age into 5 year long groups to make it easier to interpret by small enough intervals that it represents the data effectively.
We find that, in general, until about 85, we see a positive correlation between the age of patient and the length of their stay, as I predicted.
We also see that patients 35+ with a chronic condition tend to stay longer in hospital than those who don't have a chronic condition. I believe this is because they need to be more closely monitored incase the condition they were admitted for affects their chronic condition, or that they were admitted for their chronic condition worsening, which might take a while to figure out the cause of the deterioration.

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]**

Dear Utopia Health department,
I have taken a look at the data you sent me, firstly I want to say that your categorising of conditions at the time of the call seems to be proving very effective. Where one might expect longer waits for the ambulance to arrive resulting in higher risk of death and longer stays in hospital, we don't see this for most categories; the cases with higher risk of death and longer stays in hospital are being prioritised and attended to faster. However, in category two cases we see that longer waits for the ambulance to arrive do result in higher risk of death and longer stays in hospital, therefore, I would advise to prioritise category two cases more, where possible. I see that the most calls to the ambulance service are made between 23:00-24:00, therefore I would suggest increasing ambulance crew on duty at this time of day.

I also found that when patients age 35+ with chronic conditions get admitted, they tend to need to stay in hospital longer than those who don't have a chronic condition so I'd suggest taking this into account where possible when assessing the category of condition given at the time of the call. In general, we see the longer the patient stays in hospital, the lower the probability of discharge. For the more serious categories, however, we see a much sharper decline after around 23 days in hospital. For the less serious categories, we rarely see patients stay past 25 days, this fact may be helpful when figuring out how many beds you need free at any one time.