MA20277 Coursework 1

Callum Gregory

```
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]**

```
# Loading and cleaning the Data
Orchids_raw <- read.csv("Orchids.csv")
Orchids <- rename(Orchids_raw, Fertilizer_Phosphate_Level=Phos,
                  Fertilizer_Potassium_Level = Potas, Fertilizer_Sulphur_Level = Sulf,
                  Planting_Date = Planting, Temperature = Temp) %>%
  mutate(Planting_Date = as_date(Planting_Date, format = "%Y-%m-%d"))
# Setting missing values to NA
Orchids$Fertilizer_Phosphate_Level[Orchids$Fertilizer_Phosphate_Level == 0] <- NA
Orchids$Fertilizer_Potassium_Level[Orchids$Fertilizer_Potassium_Level == 0] <- NA
Orchids$Fertilizer_Sulphur_Level[Orchids$Fertilizer_Sulphur_Level == 0] <- NA
# Putting the Orchids data into tidy data format for ease later on
Orchids_tidy <- Orchids %>% rename(Phosphate = Fertilizer_Phosphate_Level,
                                   Potassium = Fertilizer_Potassium_Level,
                                   Sulphur = Fertilizer_Sulphur_Level) %>%
  pivot_longer(cols = Phosphate:Sulphur, names_to = "Chemical_in_Fertilizer") %>%
  rename(Chemical_Level = value)

# Extracting and providing the first two rows of the data set
Orchids %>% slice_head(n=2)
```

```
  Height Fertilizer_Phosphate_Level Fertilizer_Potassium_Level
1   16.3                        89                         270
2    2.6                        NA                         265
  Fertilizer_Sulphur_Level Planting_Date          Type Temperature Quality
1                       38    2022-03-19  Phalaenopsis        27.7       7
2                       39    2022-04-01 Odontoglossum        18.1       5
```

```
# Stating the minimum and maximum observed phosphate, potassium and sulphur levels
Orchids_tidy %>% filter(Chemical_Level!=0) %>% group_by(Chemical_in_Fertilizer) %>%
  summarise("Maximum Level (ppm)" = max(Chemical_Level),
            "Minimum Level (ppm)" = min(Chemical_Level))
```

```
# A tibble: 3 x 3
  Chemical_in_Fertilizer `Maximum Level (ppm)` `Minimum Level (ppm)`
  <chr>                                 <int>                 <int>
1 Phosphate                               130                    46
2 Potassium                               385                   195
3 Sulphur                                  46                    28
```

- The maximum phosphate fertilizer level is 130ppm and the minimum is 46ppm.
- The maximum potassium fertilizer level is 385ppm and the minimum is 195ppm.
- The maximum sulphur fertilizer level is 46ppm and the minimum is 28ppm.

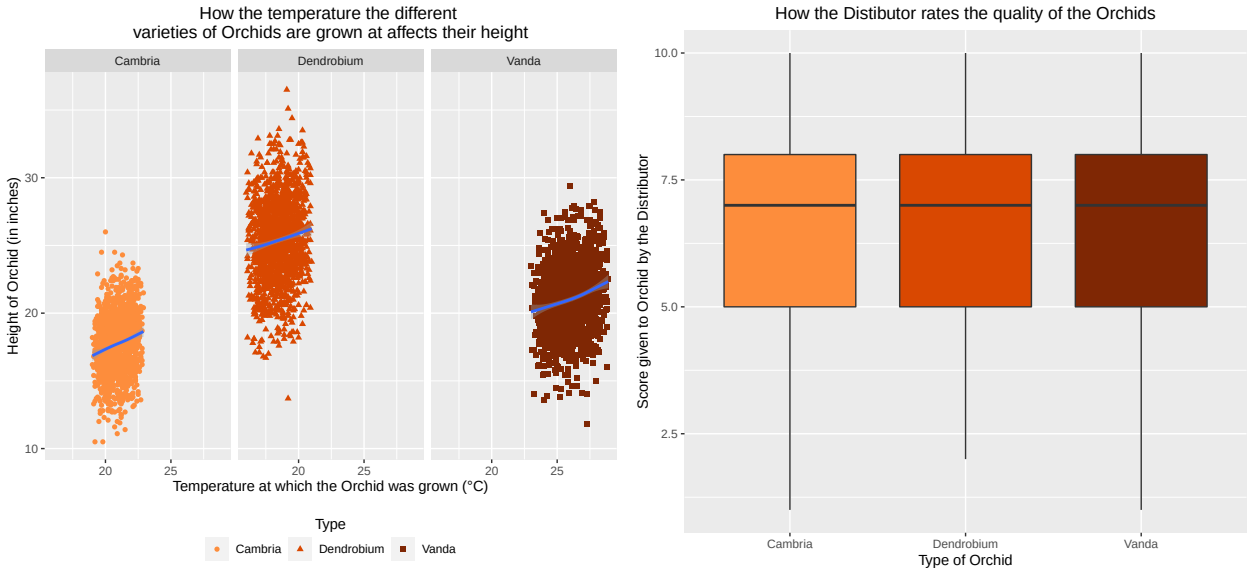
b) *Explore the relationship of temperature and plant height for the three types of orchid with the highest average height. Further investigate how these three types compare regarding their quality.* **[5 marks]**

```
# Finding the 3 types of Orchid with the highest average height
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
```

```
# Making a plot to explore the relationship between temperature and plant height
my_palette <- RColorBrewer::brewer.pal(9, "Oranges")[c(5,7,9)]
ggplot(Orchids %>% filter(Type %in% c("Dendrobium", "Vanda", "Cambria")),
       aes(x = Temperature, y = Height)) + facet_wrap(~Type) +
  geom_point(aes(shape = Type, color = Type)) + geom_smooth() +
  labs(x = "Temperature at which the Orchid was grown (\u00B0C)",
       y = "Height of Orchid (in inches)", title = "How the temperature the different
varieties of Orchids are grown at affects their height") + scale_color_manual(values =
my_palette) + theme(plot.title = element_text(hjust = 0.5), legend.position = "bottom") +
  guides(color = guide_legend(title.position = "top", title.hjust = 0.5),
         shape = guide_legend(title.position = "top", title.hjust = 0.5))
# Investigating how the 3 types of Orchid compare regarding their quality
ggplot(Orchids %>% filter(Type %in% c("Dendrobium", "Vanda", "Cambria")),
```

```
      aes(x = Type, y = Quality)) + geom_boxplot(aes(fill=Type)) +
  labs(x = "Type of Orchid", y = "Score given to Orchid by the Distributor",
       title = "How the Distibutor rates the quality of the Orchids") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none") +
  scale_fill_manual(values = my_palette)
```

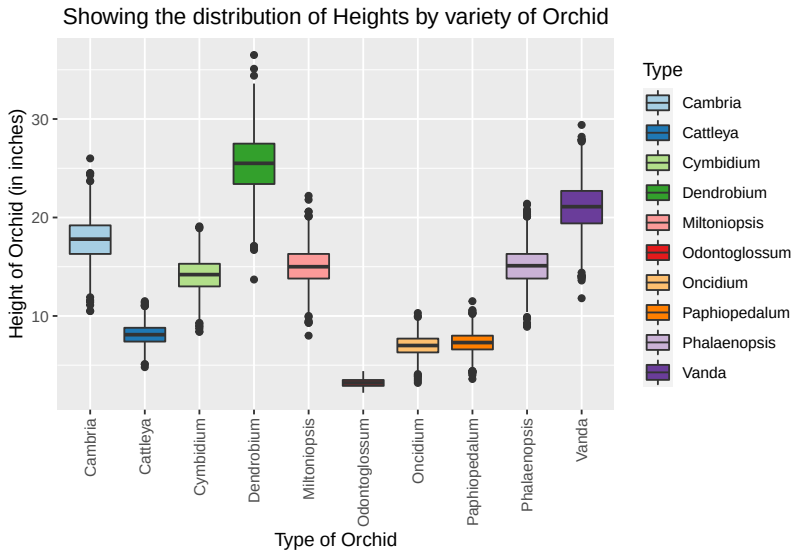


From observing the data graphic on the left, the 3 Orchids with the highest average height are grown under different ranges in temperature. More explicitly, Cambria grows between $\sim 18.5-23°C$, whereas Dendrobium grows between $\sim 16°C - 21°C$ and Vanda between $\sim 23°C - 29°C$. Overall Dendrobium Orchids grow the tallest, followed by Vanda Orchids and then Cambria Orchids and there also appears to be more variation in height of the Dendrobium Orchids compared to the other two since there is a bigger vertical spread of data points in the Dendrobium plot. Furthermore, respective to each Orchid's temperature ranges the higher the temperature the taller the Orchid in general - this is shown via `geom_smooth()` the "trend" lines.

With regards to the quality of Dendrobium, Cambria and Vanda Orchids the average quality for each is roughly the same (~ 6.5) and the distribution of the quality is very similar for all 3, so in theory equal proportions of each of these 3 Orchids should be bought by the distributor. The only real difference in the distribution of them is that the lower 25% of Dendrobium Orchids have a slightly higher quality range than that of both Cambria and Vanda.

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

```
# Investigating the differences between the types of orchids in terms of
# their distribution of height.
ggplot(Orchids, aes(x = Type, y = Height)) + geom_boxplot(aes(fill = Type)) +
  scale_fill_brewer(palette = "Paired") +
  labs(x = "Type of Orchid", y = "Height of Orchid (in inches)",
       title = "Showing the distribution of Heights by variety of Orchid") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

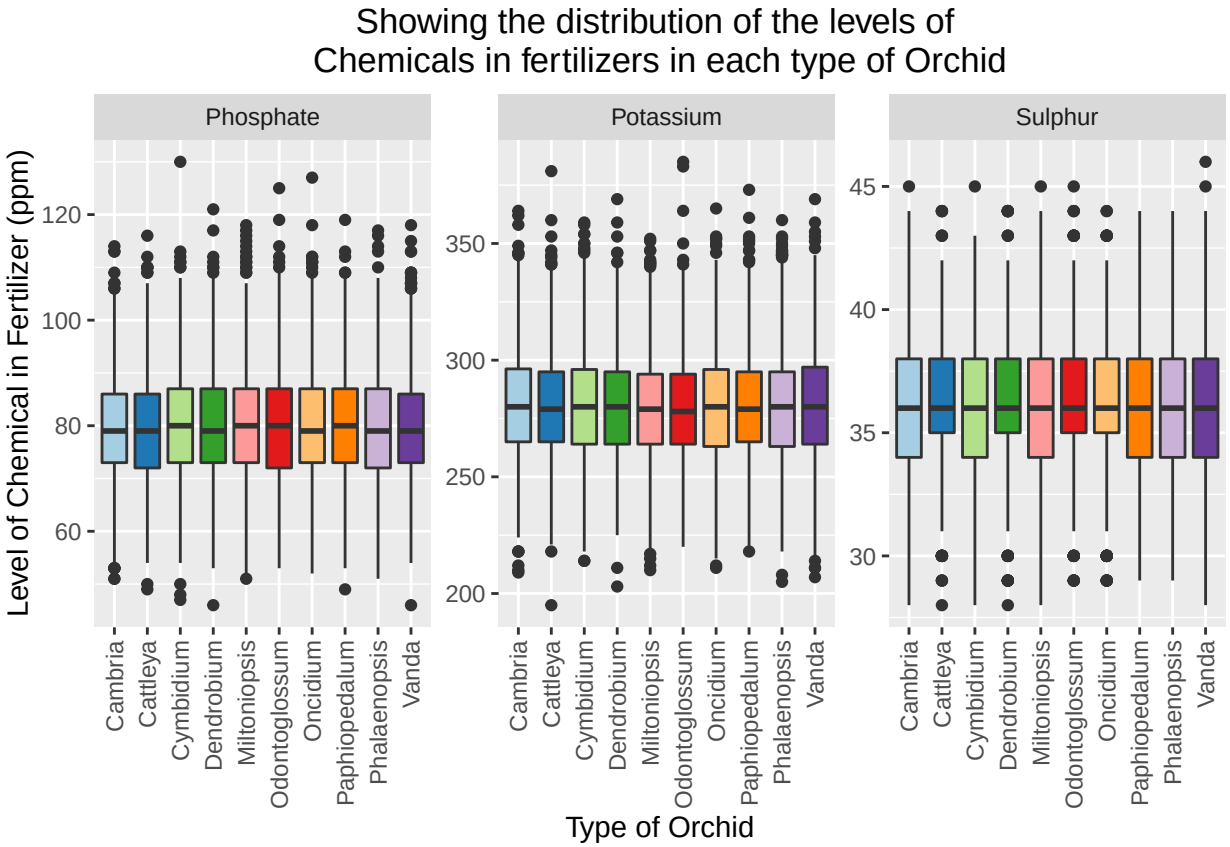


From this boxplot created, in general the greater the height of the Orchid, the more variety in its height. For example, the taller Orchids such as Dendrobium, Vanda and Cambria have a height range of about 10 to 15 inches excluding the outliers, whereas the shorter Orchids such as Odontogloddum, Oncidium and Paphiopedalum have heights that only range within about 5 inches or less.

```
# Investigating the differences in growing conditions (temperature)
Orchids %>% group_by(Type) %>% summarise(Average_Temperature=mean(Temperature))
```

```
# A tibble: 10 x 2
   Type          Average_Temperature
   <chr>                       <dbl>
 1 Cambria                      21.0
 2 Cattleya                     21.0
 3 Cymbidium                    18.5
 4 Dendrobium                   18.5
 5 Miltoniopsis                 18.5
 6 Odontoglossum                18.5
 7 Oncidium                     21.0
 8 Paphiopedalum                21.0
 9 Phalaenopsis                 26.0
10 Vanda                        26.1
```

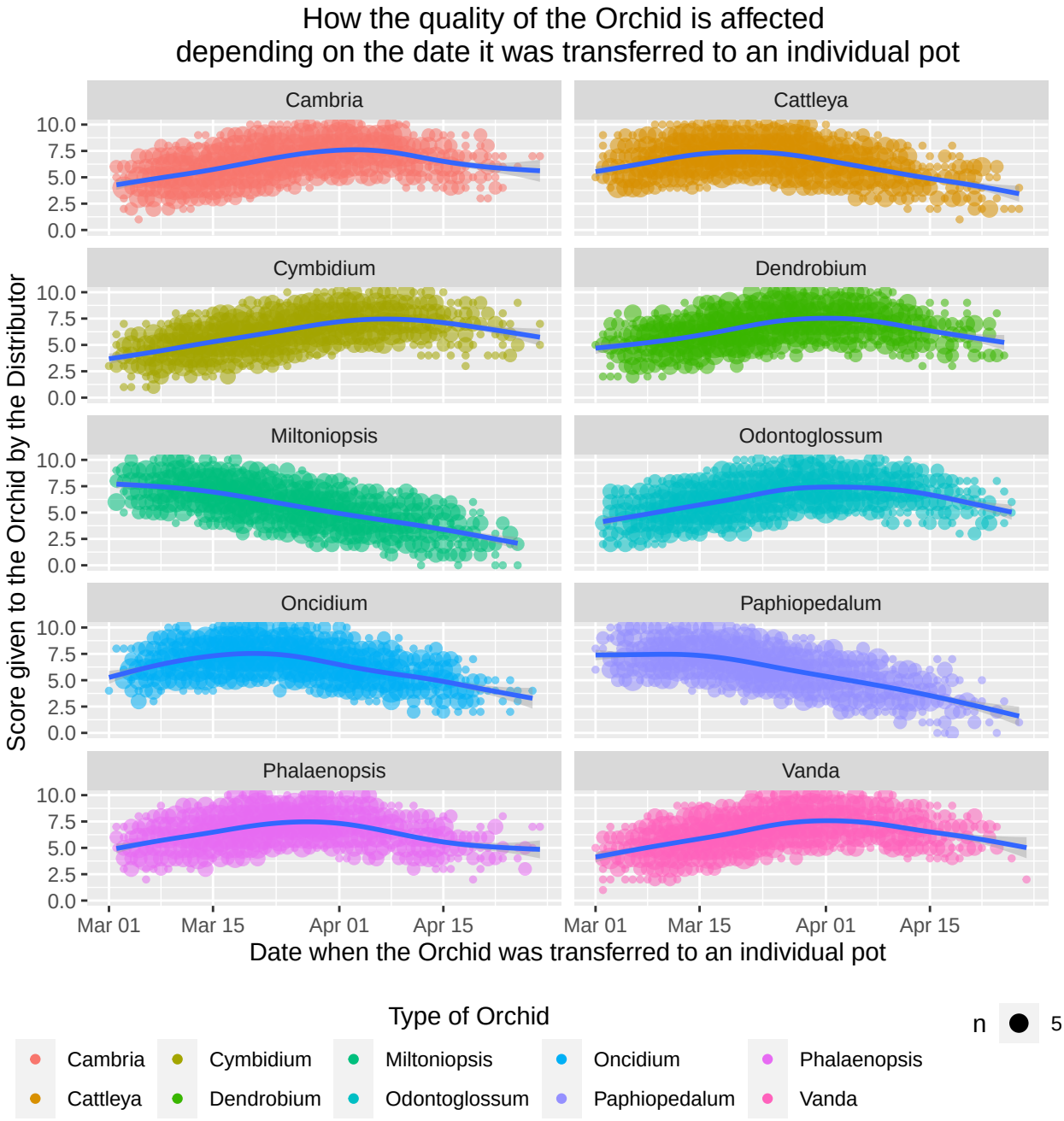
```
# Investigating the differences in growing conditions (chemical levels in fertilizer)
ggplot(Orchids_tidy %>% filter(Chemical_Level!=0), aes(x = Type, y = Chemical_Level)) +
  geom_boxplot(aes(fill = Type)) + scale_fill_brewer(palette = "Paired") +
  facet_wrap(~Chemical_in_Fertilizer, scales = "free") +
  labs(x = "Type of Orchid", y = "Level of Chemical in Fertilizer (ppm)",
       title = "Showing the distribution of the levels of
Chemicals in fertilizers in each type of Orchid", fill = "Type of Orchid") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



The different varieties of Orchids were grown at differing temperatures - see tibble - and it would be worth knowing more information about Orchids to discover if these differing temperatures were stunting the growth of some types of Orchids. By observing the boxplot, we see slight differences in the distribution of levels of each chemical in the fertilizer for differing types of Orchids (the missing values for chemical levels in fertilizers have been removed before plotting to avoid skewing the result), but in general the average chemical levels in the fertilizers appear to be fairly consistent whilst the other quartiles have slightly more variation when varying across the different types of Orchids. Another difference in growing conditions is the date when each Orchid was transferred to an individual pot which is looked at in more detail in the following question.

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**]

```
# Observing the relationship between the quality of the Orchid and the date it was
# transferred to an individual pot and how this varies based on the type of Orchid
ggplot(Orchids, aes(x = Planting_Date, y = Quality)) + facet_wrap(~Type, ncol = 2) +
  geom_count(aes(color = Type, alpha=.1)) + geom_smooth() +
  labs(x = "Date when the Orchid was transferred to an individual pot",
       y = "Score given to the Orchid by the Distributor",
       title = "How the quality of the Orchid is affected
depending on the date it was transferred to an individual pot",
       color = "Type of Orchid") + theme(plot.title = element_text(hjust = 0.5),
legend.position = "bottom") + guides(color = guide_legend(title.position = "top",
title.hjust = 0.5))
```



In 2023 the Orchid grower should plant Militoniopsis and Paphiopedalum Orchids first as both these types of Orchids grow to their best quality at the start of March rather than later on in the spring. There appears to be a little bit more flexibility on when the Orchid grower should plant the Paphiopedalum Orchid as the average quality appears the same within the first couple of weeks and for that reason, the first Orchid that the Orchid grower should plant in 2023 should be a Militoniopsis Orchid on the 1st of March. These suggestions are based on the assumption that the growing conditions in 2023 will be identical to those in 2022 - which is not realistic as there are many factors that are likely to vary e.g. temperature and levels of chemicals in fertilisers - and that the same distributor is rating the quality of the Orchids to avoid potential bias.

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]**

```
# Loading and cleaning the data
Ambulance_raw <- read.csv("Ambulance.csv")
Hospital_raw <- read.csv("Hospital.csv")
Ambulance_Service <- full_join(Ambulance_raw, Hospital_raw, by = "PatientID")
Ambulance_Service <- Ambulance_Service %>%
mutate(Call = ymd_hms(Call), Arrival = ymd_hms(Arrival), Hospital = ymd_hms(Hospital)) %>%
  rename(Initial_Response_Category = Category1, Ambulance_Arrival_Time = Arrival,
         Arrival_Response_Category = Category2, Hospital_Arrival_Time = Hospital,
         Length_of_Hospitalisation = Length)
# Finding the time of day when we tend to see the highest frequency of calls to the
# ambulance service
Calls_to_Ambulance <- Ambulance_Service[!is.na(Ambulance_Service$Call),]
Calls_to_Ambulance %>% mutate(Time_of_Day = hour(Call)) %>% group_by(Time_of_Day) %>%
  summarise(Number_of_Calls = n()) %>% arrange(desc(Number_of_Calls)) %>% slice_head(n=3)
```

```
# A tibble: 3 x 2
  Time_of_Day Number_of_Calls
        <int>           <int>
1          17            4883
2          16            4749
3           7            4448
```

By categorising calls to the ambulance service hourly, we can see that the highest frequency of calls to the ambulance service is between 17:00-18:00.

```
# Finding the proportion of calls which lead to the patient being delivered to hospital
1-sum(is.na(Calls_to_Ambulance$Hospital_Arrival_Time))/nrow(Calls_to_Ambulance)
```

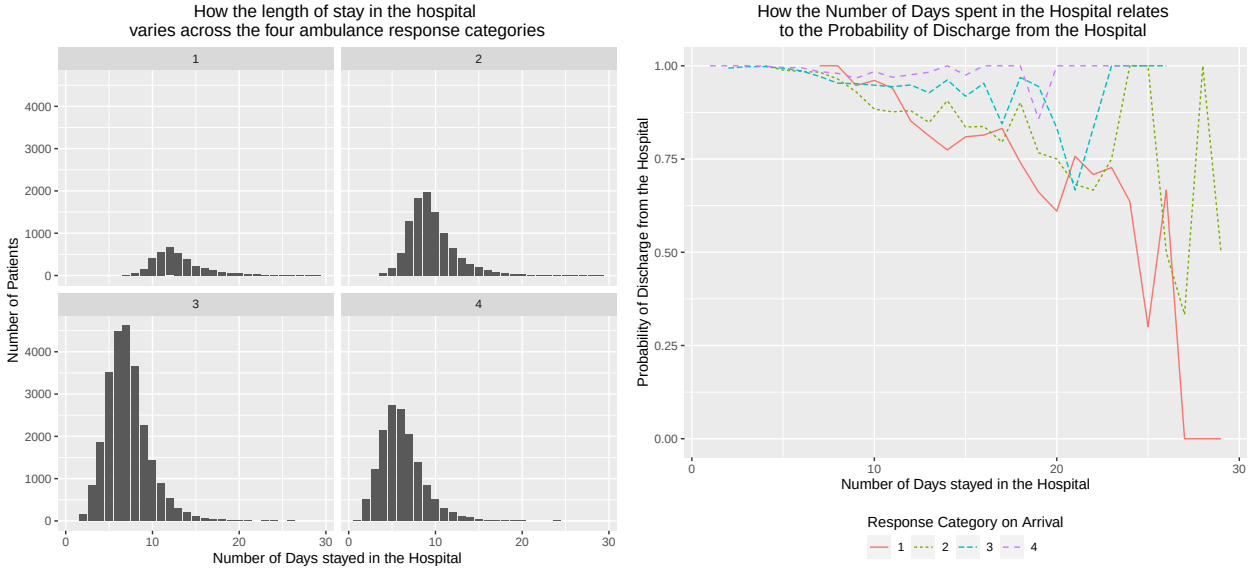
```
[1] 0.8002888
```

The proportion of calls that lead to the patient being delivered to hospital is ~ 0.8.

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]**

```
# Finding the patients admitted to the hospital via the Ambulance
Hospitalised_Patients<-Ambulance_Service[!is.na(Ambulance_Service$Hospital_Arrival_Time),]
ggplot(Hospitalised_Patients, aes(x = Length_of_Hospitalisation)) + geom_bar() +
  facet_wrap(~Arrival_Response_Category) + theme(plot.title = element_text(hjust = 0.5)) +
```

```
  labs(x = "Number of Days stayed in the Hospital", y = "Number of Patients",
       title = "How the length of stay in the hospital
varies across the four ambulance response categories")
Length_of_Stay <- Hospitalised_Patients %>%
  group_by(Arrival_Response_Category, Length_of_Hospitalisation) %>%
  summarise("Probability_of_Discharge" = sum(Outcome == "0")/(sum(Outcome == "0") +
                                                               sum(Outcome == "1")))
Length_of_Stay$Arrival_Response_Category <-
  as.character(Length_of_Stay$Arrival_Response_Category)
ggplot(Length_of_Stay, aes(x = Length_of_Hospitalisation, y = Probability_of_Discharge)) +
geom_line(aes(linetype = Arrival_Response_Category, color = Arrival_Response_Category)) +
  labs(x = "Number of Days stayed in the Hospital",
       y = "Probability of Discharge from the Hospital",
       title = "How the Number of Days spent in the Hospital relates
to the Probability of Discharge from the Hospital ",
       linetype = "Response Category on Arrival",
       color = "Response Category on Arrival") +
  theme(plot.title = element_text(hjust = 0.5),legend.position = "bottom") +
  guides(color = guide_legend(title.position = "top", title.hjust = 0.5),
         linetype = guide_legend(title.position = "top", title.hjust = 0.5))
```



From the data graphic on the left, we can conclude that the non-urgent (Category 4) patients, unsurprisingly tended to have the shortest stays in hospital and that for each more life threatening category the distribution of the number of patients staying in hospital shifts to the right, so that patients stayed in hospital for more days all round - this makes sense as more life threatening cases are going to take up more of the hospital's time due to more extensive medical testing and longer after-care procedures.

From the graph on the right the patients categorised as being in life threatening condition (Category 1) are guaranteed to spend a longer time in hospital. The earliest someone from Category 1 is discharged is at least 7 days after they first entered the hospital whereas for Category 2 it's 4 days, Category 3 it's 2 days and Category 4 it's 1 day. Hence, there is a trend that if the patient is in a lower risk category, they are likely to be discharged quicker from the hospital.

Secondly, we can see that as the number of days spent in the hospital increases, Category 4 patients remain with a very high probability of being discharged. In the other 3 Categories however, we notice the general trend that as the number of days spent in hospital increases the probability of discharge decreases. More

specifically, Category 1 decreases the most and generally has a lower probability of being discharged compared to Category 2 and 3 - something that seems likely given that Category 1 is for patients who are in life-threatening condition. We also notice that towards the right hand side of the graph where the number of days spent in the hospital is large, there are large spikes. This is due to having little data to analyse as most patients don't spend that long in hospital - this is easy to see from the graph to the left - and thus for longer stays in hospital, each individual patient will have a bigger weight on the overall probability of discharge, hence the large spikes.

```
# Finding the probability of discharge from hospital
Hospitalised_Patients %>% group_by(Arrival_Response_Category) %>%
  summarise("Probability_of_Discharge" = sum(Outcome == "0")/(sum(Outcome == "0") +
                                                               sum(Outcome == "1")))
```

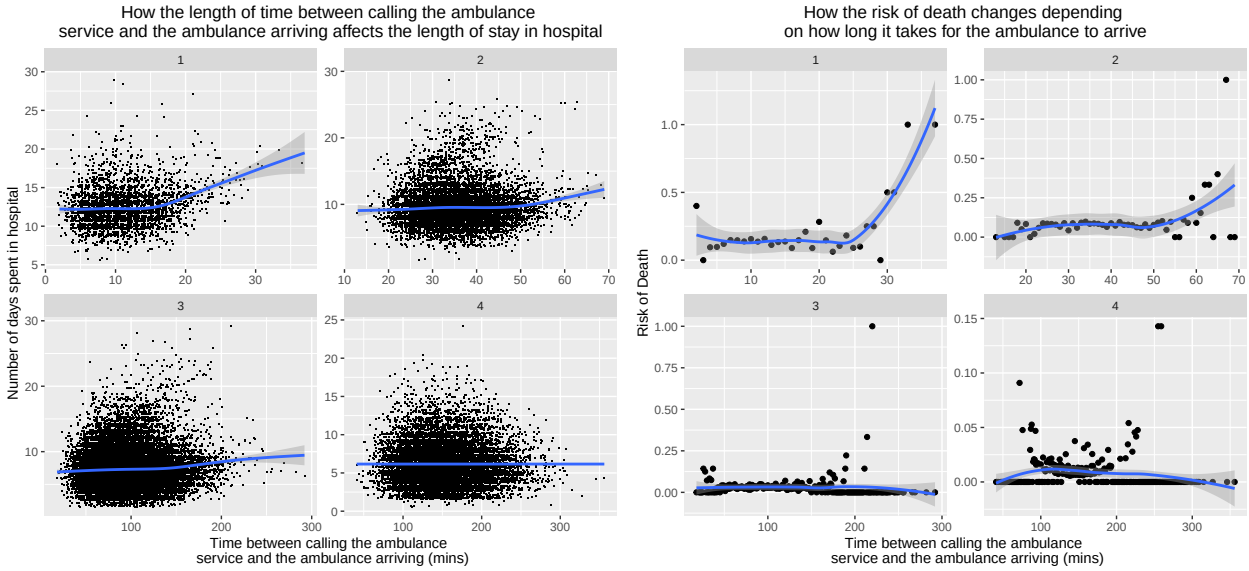
```
# A tibble: 4 x 2
  Arrival_Response_Category Probability_of_Discharge
                      <int>                    <dbl>
1                         1                    0.850
2                         2                    0.921
3                         3                    0.972
4                         4                    0.990
```

From observing the tibble we can see that the patients in life threatening condition (Category 1) have the lowest probability of being discharged with the probability of discharge increasing as each category becomes less life threatening ending with (Category 4) non-urgent patients having the highest chance of discharge from the hospital.

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]**

```
# Investigating length of stay and time taken for ambulance to arrive
Hospitalised_Patients <- Hospitalised_Patients %>%
  mutate(Time_to_reach_patient = Ambulance_Arrival_Time-Call)

ggplot(Hospitalised_Patients, aes(x=Time_to_reach_patient, y=Length_of_Hospitalisation)) +
geom_jitter(shape=".") + facet_wrap(~Initial_Response_Category, scales = "free") +
labs(x = "Time between calling the ambulance \n service and the ambulance arriving (mins)",
     y = "Number of days spent in hospital",
     title = "How the length of time between calling the ambulance
service and the ambulance arriving affects the length of stay in hospital") +
  theme(plot.title = element_text(hjust = 0.5)) + geom_smooth()
# Investigating risk of death and time taken for the ambulance to arrive
Death_Discharge <- Hospitalised_Patients %>% group_by(Initial_Response_Category,
                                                      Time_to_reach_patient) %>%
  summarise("Death_Count" = sum(Outcome == "1"), "Discharge_Count" = sum(Outcome == "0"),
            "Risk_of_Death" = Death_Count/(Death_Count + Discharge_Count))
Death_Discharge$Initial_Response_Category <-
  as.character(Death_Discharge$Initial_Response_Category)
ggplot(Death_Discharge, aes(x = Time_to_reach_patient, y = Risk_of_Death)) +
geom_point() + facet_wrap(~Initial_Response_Category, scales = "free") + geom_smooth() +
labs(x = "Time between calling the ambulance \n service and the ambulance arriving (mins)",
     y = "Risk of Death", title = "How the risk of death changes depending
on how long it takes for the ambulance to arrive") +
  theme(plot.title = element_text(hjust = 0.5))
```



We have seen in the previous question that the response category the ambulance puts the patients in when it arrives has an impact on the length of the patient's stay in hospital and hence it makes sense to assume that the initial response category may have an impact in the speed at which the ambulance arrives, so we split the data into these categories.

For category 1 and 2 patients, the ambulance arrives to every patient within ~ 50 and ~ 70 mins of the call respectively, whereas for the lower risk categories (3 and 4) there is a much bigger range of time in which the ambulance will take to arrive after receiving the phone call. In terms of the length of the patients stay in hospital, for category 1 patients after ~ 20 mins the length of stay begins to increase very quickly, whereas before ~ 20 mins it averages at about 12 to 13 days. For category 2 and 3 patients the average number of days staying in the hospital remains roughly the same until after ~ 50 mins and ~ 150 mins respectively when it begins to increase at a slow rate. On the contrary, category 4 patients average around 6 to 7 days in hospital regardless of the time it took for the ambulance to arrive after the call.

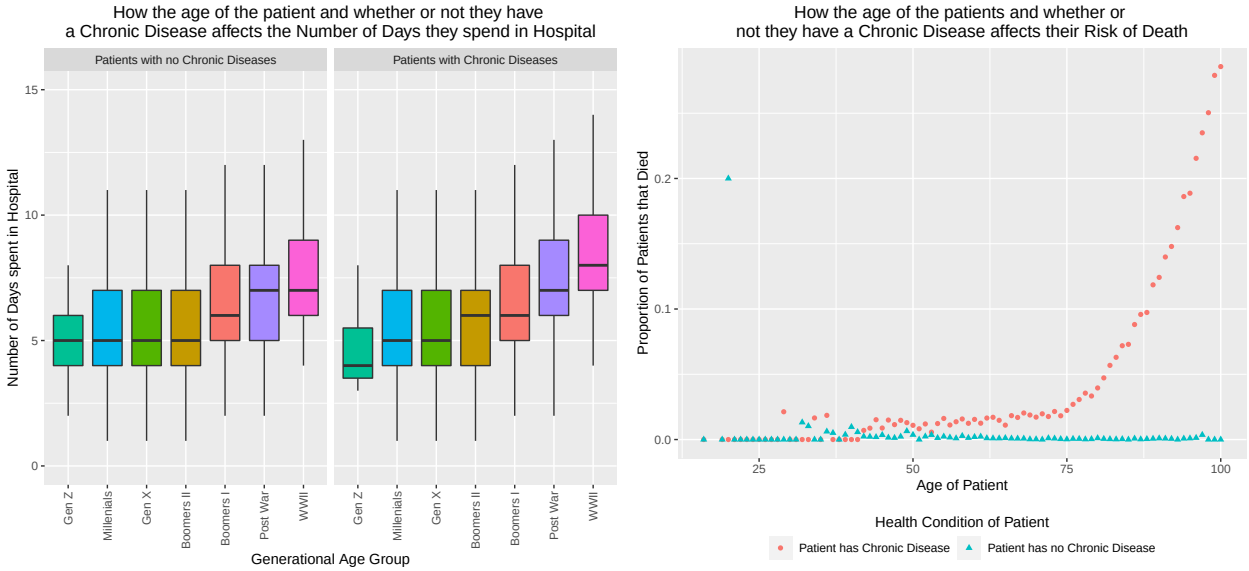
For category 1 patients, the risk of death hovers at $\sim 20\%$ if the ambulance arrives within 25 minutes of the call but as soon as the ambulance takes longer, the risk of death increases quickly. For category 2 patients, if the ambulance arrives within 25 minutes of the call, the risk of death is very low but from 25 minutes and onwards the risk of death hovers at $\sim< 10\%$ until the ambulance arrives ~ 50 minutes after the call where the risk of dying then begins to slowly increase approaching $\sim 30\%$ once the ambulance takes ~ 70 minutes to arrive. For patients in both categories 3 and 4 we see a very low risk of death regardless of when the ambulance arrives after the patient's call.

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**]

My Question: How does the age of patients and whether or not they have any Chronic Diseases affect the length of stay in Utopia Hospital and their outcome?

```
Hospital_data <- Ambulance_Service[!is.na(Ambulance_Service$Length_of_Hospitalisation),]
# Observing how the age of the patient and whether they have Chronic Diseases or not
# affects the length of their stay in hospital
Hospital_data$Chronic <- as.character(Hospital_data$Chronic)
Hospital_data <- Hospital_data %>% mutate(Age_Group = case_when(Age > 9 & Age <= 25 ~
"Gen Z", Age > 25 & Age <= 41 ~ "Millenials", Age > 41 & Age <= 57 ~ "Gen X", Age > 57 &
Age <= 67 ~ "Boomers II", Age > 67 & Age <= 76 ~ "Boomers I", Age > 76 & Age <= 94 ~
"Post War", Age > 94 & Age <= 100 ~ "WWII" ))
```

```
ggplot(Hospital_data, aes(x = factor(Age_Group, level = c("Gen Z", "Millenials", "Gen X",
"Boomers II", "Boomers I", "Post War", "WWII" )), y = Length_of_Hospitalisation)) +
  geom_boxplot(aes(fill = Age_Group), outlier.shape = NA) + facet_wrap(~Chronic,
labeller = as_labeller(c(`0`="Patients with no Chronic Diseases",
`1`="Patients with Chronic Diseases"))) + labs(x = "Generational Age Group", title =
"How the age of the patient and whether or not they have
a Chronic Disease affects the Number of Days they spend in Hospital",
y = "Number of Days spent in Hospital") + theme(legend.position="none", plot.title =
element_text(hjust = 0.5), axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust = 1)) + ylim(0,15) # Have removed outliers to better see the distribution of data
# Observing how the age of the patient and whether they have Chronic Diseases or not
# affects the risk of death
Proportions <-Hospital_data %>% group_by(Age) %>%
  summarise("Chronic_Disease" = sum(Outcome == "1"& Chronic == "1")/(sum(Chronic == "1") +
            sum(Chronic == "0")), "Non_Chronic_Disease"= sum(Outcome == "1" &
            Chronic == "0")/(sum(Chronic == "1") + sum(Chronic == "0")))
Proportions_tidy <- Proportions %>%
  pivot_longer(cols = Chronic_Disease:Non_Chronic_Disease,
               names_to = "Health_Condition") %>% rename(Proportion_of_Deaths = value)
ggplot(Proportions_tidy, aes(x = Age, y = Proportion_of_Deaths)) +
  geom_point(aes(color = Health_Condition, shape = Health_Condition)) +
  labs(x = "Age of Patient", y = "Proportion of Patients that Died",
       title = "How the age of the patients and whether or
not they have a Chronic Disease affects their Risk of Death",
color = "Health Condition of Patient", shape = "Health Condition of Patient") +
  scale_color_discrete(labels = c("Patient has Chronic Disease",
                                  "Patient has no Chronic Disease")) +
  scale_shape_discrete(labels = c("Patient has Chronic Disease",
                                  "Patient has no Chronic Disease")) +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "bottom") +
  guides(colour = guide_legend(title.position = "top", title.hjust = 0.5),
         shape = guide_legend(title.position = "top", title.hjust = 0.5))
```



From the data graphic on the left we have used the classification from Beresford Research[1] to categorise the age of patients into the corresponding Generational Groups. On splitting into generational age categories

it is noticeable that the whole distribution of the Millenial, Gen X and Boomers II age groups are similar regardless of whether the patient has a Chronic Disease or not. However, more explicitly we see that the mean number of days spent in hospital for the Boomers II generation is greater if the patient does have a Chronic Disease. Furthermore, if the patient is older than 76yrs (so in the Boomers I, Post War or WWII generation) the length of stay in hospital will increase across these 3 generations (so that the whole distribution of WWII is over the largest number of days in hospital) and this happens regardless of whether the patient has a Chronic Disease or not. For Gen Z, we see a slightly different trend and we see that the mean number of days spent in hospital is actually lower if you have a Chronic Disease, but it is worth mentioning how few patients there in Gen Z - see tibble below - so the information regarding this category is less reliable.

```
Hospital_data %>% group_by(Age_Group, Chronic) %>% summarise("Total Number of Patients" =
sum(Chronic==0)+sum(Chronic==1))%>%filter(Age_Group=="Gen Z")
```

```
# A tibble: 2 x 3
# Groups:   Age_Group [1]
  Age_Group Chronic `Total Number of Patients`
  <chr>     <chr>                         <int>
1 Gen Z     0                                53
2 Gen Z     1                                 7
```

The data graphic on the right highlights how for all ages, the risk of death is very low if the patient has no Chronic Diseases. However, it is to be noted that there is a risk of death quite high if the patient is 20yrs old, but it's probably safe to assume this is an outlier given that there are only 5 patients who are 20yrs old in the data (so this one death has had a big impact on the probability of death). We also see that for patients with Chronic Diseases, the risk of death is very low for all ages until you reach $\sim$ 75yrs. Once the patient is over 75 years old the risk of death begins to increase in what appears to be an exponential manner.

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 investigating the data provided by Utopia's health department regarding their hospital and ambulance service it can be concluded that the Ambulance Service was busiest between 16:00-18:00 in the evening and 07:00-08:00 in the morning, and hence it'd be worth Utopia's ambulance service having more staff employed at these times to help cope with the demand (supposing they have plentiful resources). It was also discovered that the ambulance service provided is efficient to responding to calls from people in need and that the ambulance service responds promptly to these patients in an orderly fashion, depending on how serious the condition of the patient is. These two factors - how the ambulance viewed the condition of the patient and how fast the ambulance responded - have ensured that the risk of death is kept low and the probability of discharge is relatively high.

We have seen that the length of hospitalisation for the patients is still fairly long for patients in the non-urgent response category, so Utopia could potentially improve their service by speeding up the flow of non-urgent patients - this would also then give more space and resources to help the patients in more urgent need. This could potentially be resolved by having split units for differing response categories. It's also seen that for most ages of patients that regardless of whether or not they have a Chronic Disease they have a very low risk of death - highlighting how the hospital is good at adhering to those with Chronic Diseases. However, for elderly patients over 75yrs having a Chronic Disease does increase their risk of death, so it appears that the Chronic Diseases put the elderly at a higher risk of death than it does to younger patients.

[1] - 2022 | *Beresford Research* | https://www.beresfordresearch.com/age-range-by-generation/