MA20277 Coursework 1

Jacob Forsdyke

```
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 sulfur levels.* **[4 marks]**

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

orchids<-rename(orchids,Phosphate=`Phos`, Potassium=`Potas`,Sulfur=`Sulf`, Temperature=`Temp`)
orchids<-filter(orchids,Phosphate!=0)
orchids<-filter(orchids,Sulfur!=0)
orchids<-filter(orchids,Potassium!=0)
head(orchids,2)
```

```
##   Height Phosphate Potassium Sulfur   Planting         Type Temperature Quality
## 1   16.3        89       270     38 2022-03-19 Phalaenopsis        27.7       7
## 2   22.7        83       295     34 2022-04-14   Dendrobium        17.5       6
```

```
orchids %>% summarise("maximum phoshate"=max(`Phosphate`),"minimum phosphate"=min(`Phosphate`),
                      "maximum potassium"=max(`Potassium`),"minimum potassium"=min(`Potassium`),
                      "maximum sulfur"=max(`Sulfur`),"minimum sulfur"=min(`Sulfur`))
```

```
##   maximum phoshate minimum phosphate maximum potassium minimum potassium
## 1              130                46               381               195
##   maximum sulfur minimum sulfur
## 1             46             28
```

I have provided the Maximum and Minimum potassium, sulfur and phosphate levels of the entire data set and not just the first two observed values, this being provided in the summary. An observation to make is that the maximum potassium levels are much higher then the other fertilizers, suggesting it is a better fertilizer.

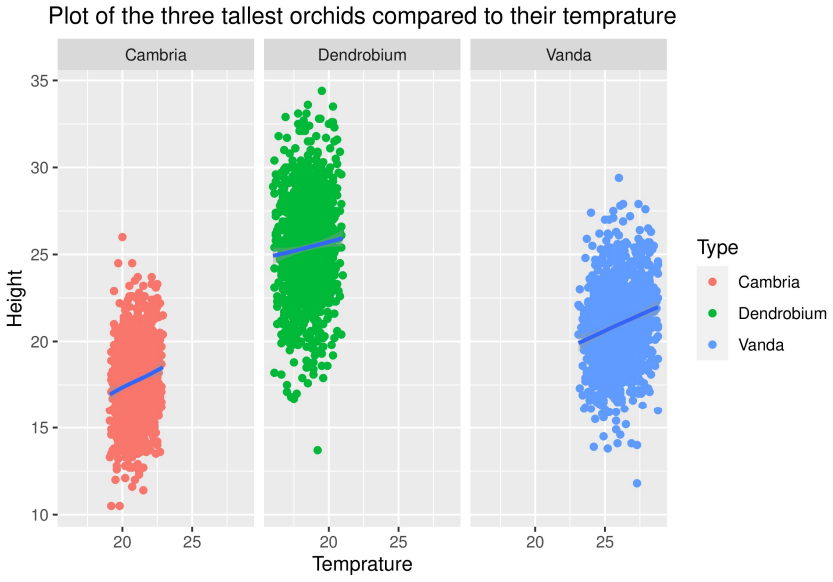
b) *Explore the relationship of temperature and plant height for the three types of orchid with the highest average height. Further investigate how these three types compare regarding their quality.* **[5 marks]**

```
orchids%>% group_by(Type) %>% summarise( "Mean height"= mean(Height))%>%
  arrange(desc(`Mean height`))
```

```
## # A tibble: 10 x 2
##    Type          `Mean height`
##    <chr>                 <dbl>
##  1 Dendrobium             25.4
##  2 Vanda                  21.0
##  3 Cambria                17.7
##  4 Phalaenopsis           15.0
##  5 Miltoniopsis           15.0
##  6 Cymbidium              14.2
##  7 Cattleya                8.09
##  8 Paphiopedalum           7.33
##  9 Oncidium                6.98
## 10 Odontoglossum           3.20
```

```
best_orchids<-filter(orchids,Type %in% c("Dendrobium", "Vanda", "Cambria"))

ggplot(best_orchids, aes(x=Temperature,y=Height)) + facet_wrap(~Type) +
  geom_point(aes(linetype=Type,color=Type)) +
  labs(title= "Plot of the three tallest orchids compared to their temprature",
       x="Temprature",
       y= "Height")+geom_smooth()+
  theme(plot.title = element_text(hjust = 0.5))
```

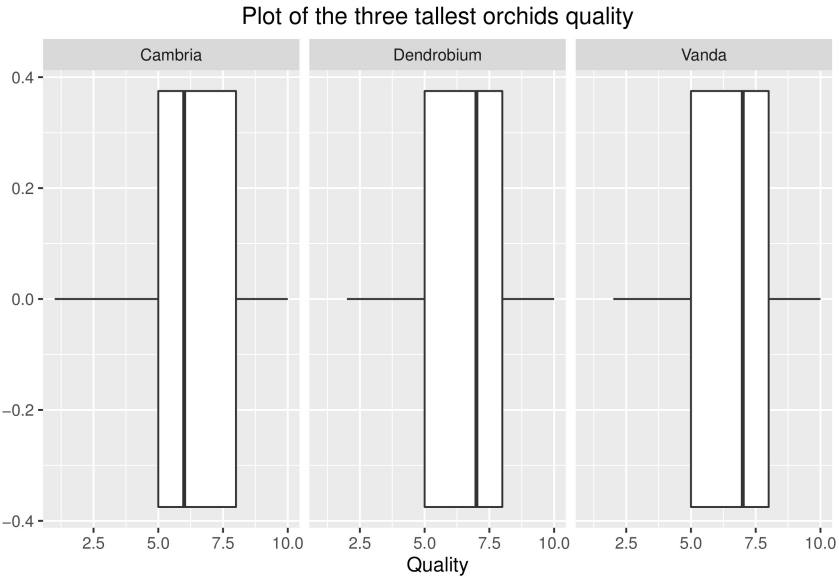


```
best_orchids%>% group_by(Type) %>%
  summarise( "mean"= mean(Quality)) %>% arrange(desc(mean))
```

```
## # A tibble: 3 x 2
##   Type        mean
##   <chr>      <dbl>
## 1 Dendrobium  6.61
## 2 Vanda       6.53
## 3 Cambria     6.48
```

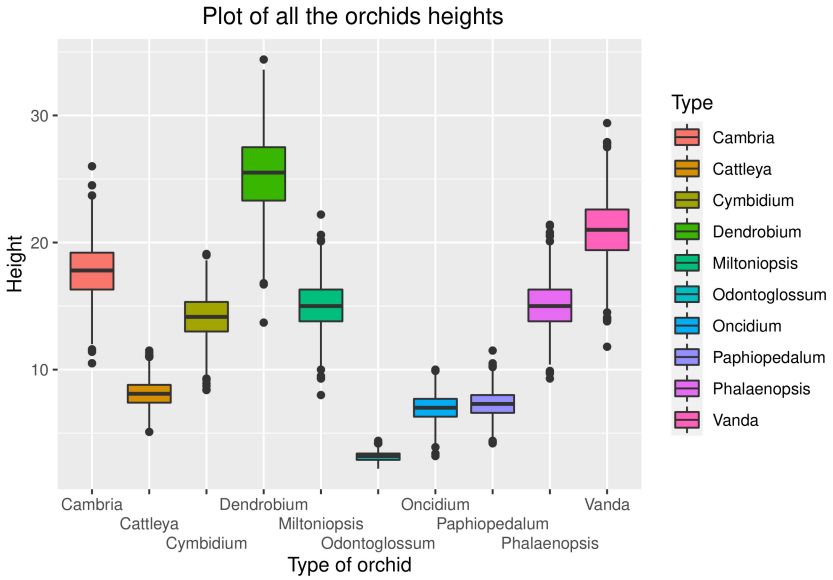
```
best_orchids<-filter(orchids,Type %in% c("Dendrobium", "Vanda", "Cambria"))

ggplot(best_orchids,aes(Quality)) + geom_boxplot() + facet_wrap(~Type)+
  labs(title= "Plot of the three tallest orchids quality")+
  theme(plot.title = element_text(hjust = 0.5))
```



As you can see as temperature increases for the three tallest average orchids the height also increases, however this is misleading due to them growing at different temperatures, dendridium doesn't grow at the temperature vanda will and yet its average height is greater. Using a boxplot i also found that the quality of each of the three orchids is nearly identical. 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( aes(fill=Type) ) +
  scale_x_discrete(guide = guide_axis(n.dodge=3))+
  labs(title=" Plot of all the orchids heights",x="Type of orchid", y="Height")+
  theme(plot.title = element_text(hjust = 0.5))
```



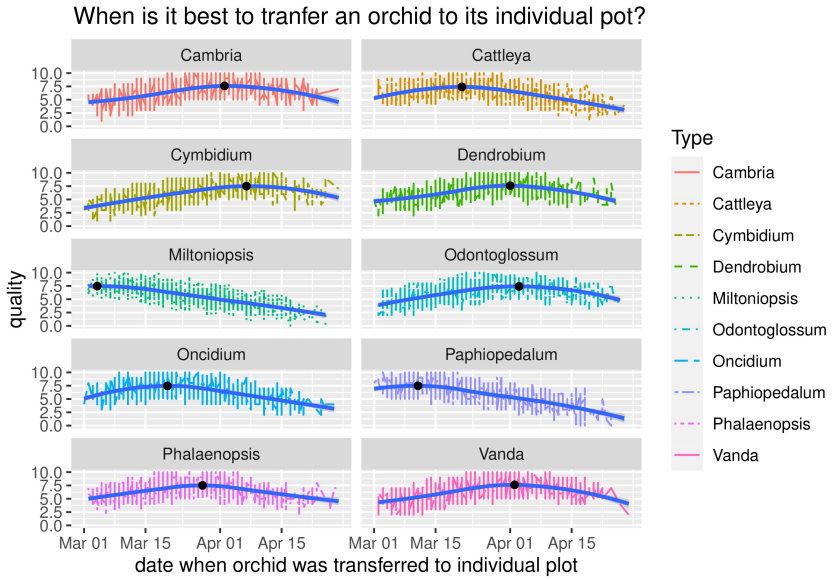

```
orchids%>% group_by(Type) %>% summarise( "Mean Potassium levels"= mean(`Potassium`),
                                         " Mean sulfur levels"= mean(`Sulfur`),
                                         "Mean phosphate levels" = mean(`Phosphate`),
                                         "Mean temprature"=mean(`Temperature`),
                                         "Mean Height"= mean(`Height`)) %>% arrange(desc(`Mean Height`))
```

```
## # A tibble: 10 x 6
##    Type          `Mean Potassium levels`  Mean sulfur ~1 Mean ~2 Mean ~3 Mean ~4
##    <chr>                           <dbl>           <dbl>   <dbl>   <dbl>   <dbl>
##  1 Dendrobium                       280.            36.2    79.9    18.5   25.4
##  2 Vanda                            281.            36.3    79.5    26.1   21.0
##  3 Cambria                          280.            36.1    79.8    21.0   17.7
##  4 Phalaenopsis                     279.            36.2    79.5    26.0   15.0
##  5 Miltoniopsis                     279.            36.3    80.6    18.5   15.0
##  6 Cymbidium                        280.            36.2    80.0    18.5   14.2
##  7 Cattleya                         280.            36.3    79.9    21.0    8.09
##  8 Paphiopedalum                    281.            36.2    80.0    21.0    7.33
##  9 Oncidium                         279.            36.3    80.0    21.1    6.98
## 10 Odontoglossum                    279.            36.4    80.0    18.5    3.20
## # ... with abbreviated variable names 1: ` Mean sulfur levels`,
## #   2: `Mean phosphate levels`, 3: `Mean temprature`, 4: `Mean Height`
```

Ordering the Heights of the orchids into descending order allows me to see if there is any correlation between height and growing conditions however this doesn't seem to be the case with the shortest orchid having higher levels than the tallest in all conditions. This may suggest that the height is effected by something not placed within the data such as light levels and water levels which is not included in the data set.

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**]

```
sm_max = orchids %>% group_by(Type) %>%
  mutate(smooth =predict(loess(Quality~as.numeric(Planting)))) %>%
  slice_max(order_by = smooth)
ggplot(orchids, aes(x=Planting,y=Quality)) + geom_line(aes(linetype=Type,color=Type)) +
  geom_smooth(method='loess')+
  labs(title=" When is it best to tranfer an orchid to its individual pot?",
                              x="date when orchid was transferred to individual plot",
                              y= "quality") + facet_wrap(~Type,ncol=2)+
  geom_point(data=sm_max,aes(y=smooth))+
  theme(plot.title = element_text(hjust = 0.5))
```




```
date<- c("2022-03-01","2022-03-02","2022-03-03","2022-03-04","2022-03-05","2022-03-06","2022-03-07")
First_week <-filter(orchids, Planting== date)
First_week%>% group_by(Type) %>% summarise( "Mean quality"= mean(Quality))%>%
  arrange(desc(`Mean quality`))
```

```
## # A tibble: 10 x 2
##    Type          `Mean quality`
##    <chr>                  <dbl>
##  1 Paphiopedalum           8
##  2 Miltoniopsis            6.71
##  3 Oncidium                6.67
##  4 Cattleya                6.09
##  5 Phalaenopsis            4.82
##  6 Cambria                 4.6
##  7 Cymbidium               4.54
##  8 Vanda                   4.5
##  9 Dendrobium              4.44
## 10 Odontoglossum           3.83
```

From the graphs you can see that the further to the left the point is the sooner the orchids should be planted,the two orchids that should be planted are miltoniopsis and paphiopedalum, this is also evident by the fact that in the first week of planting their mean quality is higher than the rest. I tried placing the dates of each of the maximums but this obscured the graphs.

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:

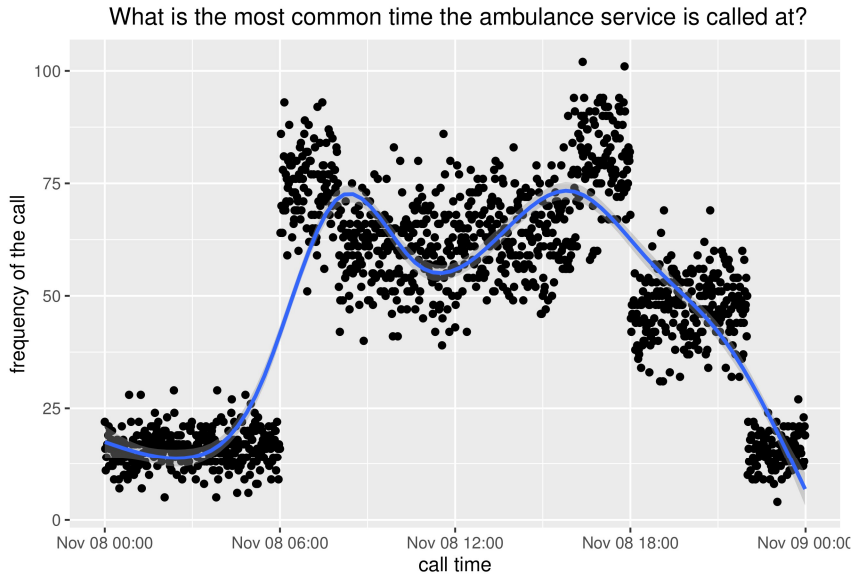
```
hospital<- read.csv( "Hospital.csv" )

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

This is the tidying up of data which will allow me to take times and date away from each other ect.

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]**

```
 Time<- format(as.POSIXct(ambulance$Call), format = "%H:%M:%S")
call_time<-ambulance %>%
  mutate( Time=format(as.POSIXct(ambulance$Call), format = "%H:%M:%S"))%>%
  group_by(Time) %>%
  summarise("freq"=n())
call_time$Time<-as.POSIXct(call_time$Time, format = "%H:%M:%S")
ggplot(call_time, aes(x=Time,y =freq)) + geom_point() +
  labs(title= " What is the most common time the ambulance service is called at?",
       x="call time",
       y= "frequency of the call") + geom_smooth()+
  theme(plot.title = element_text(hjust = 0.5))
```



```
(colMeans(is.na(ambulance)))*100
```

```
##       Call Category1   Arrival Category2 PatientID  Hospital
##    0.00000   0.00000   0.00000   0.00000   0.00000  19.97112
```

As you can see from the graph the most frequent times that the ambulance is called is during the day when people are awake, with the peak being at around 8 am and 5 pm this would likely be due to rush hour being at these times increasing the likelihood of an accident. The proportion of call that lead to a patient being delivered to the hospital is also around 20 percent.

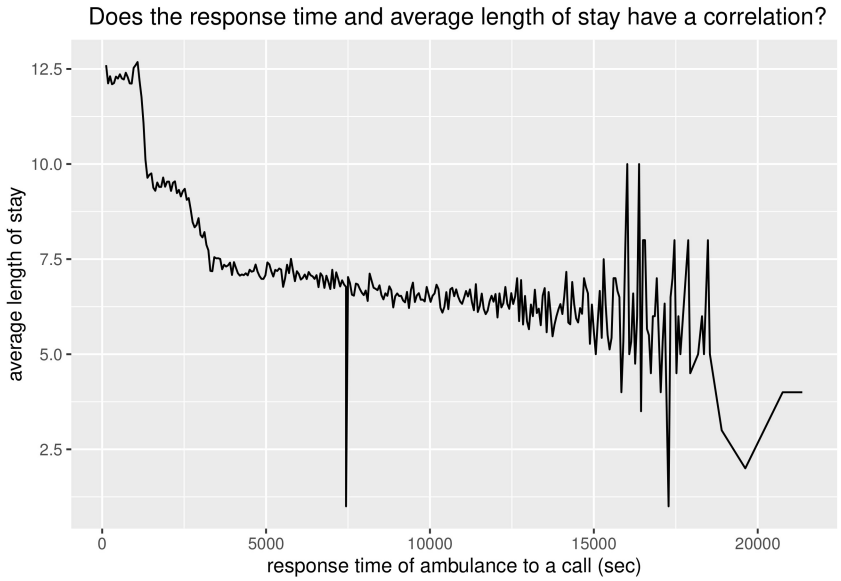
b) *How does the length of stay in hospital and the probability of discharge from hospital vary across the four ambulance response categories? Here, ambulance response category refers to that at the time of arrival of the ambulance.* [**4 marks**]

```
ambulance_hospital<-left_join(hospital, ambulance, by= "PatientID")
ambulance_hospital<-ambulance_hospital[!is.na(ambulance_hospital$Category2),]
ambulance_hospital<-ambulance_hospital %>%
  group_by(Length)%>%
  mutate('Prob of discharge'=sum(Outcome==0)/(sum(Outcome==0) +sum(Outcome==1)))
ggplot(ambulance_hospital,aes(x=Length, y=`Prob of discharge`)) + geom_line(size=2)+
  facet_wrap(~Category2) +
  labs(title= "Plot comparing length of stay to probability of being discharged",
       x="Length of time the patient spent in the hospital(days)",
       y= " the probability of the patient being discharged")+
  theme(plot.title = element_text(hjust = 0.5))
```

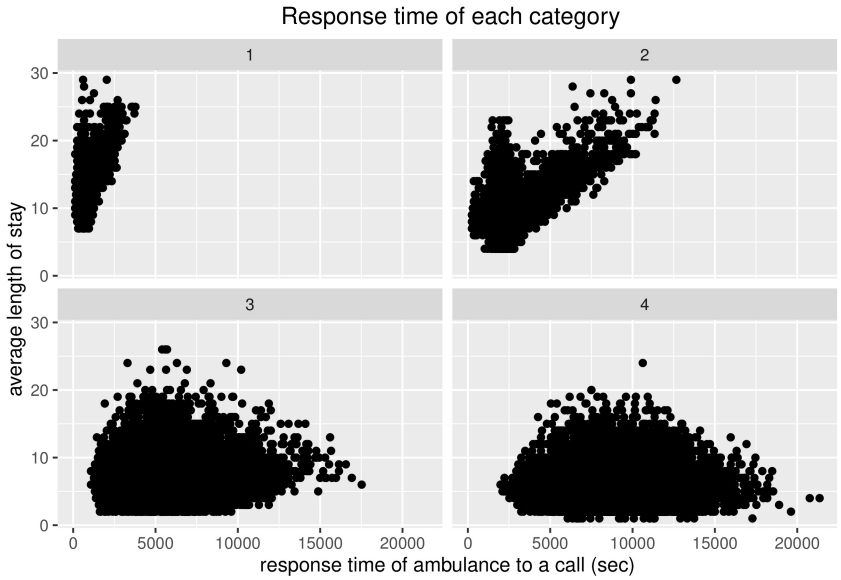
The graphs indicate that across all category the longer the patient stays in hospital the more likely they are less likely they are to be discharged, with the categories 1 and 2 having the most severe drop in chance, this is to be expected as these are the most injured patients which makes them more likely to die, however i would expect the trend to be that the longer a patient stay the less likely they would be to die. This is because i would expect the majority of patients to be at hospital for injuries and the longer they stay the more there wounds would be treated and they would start to heal.

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**]

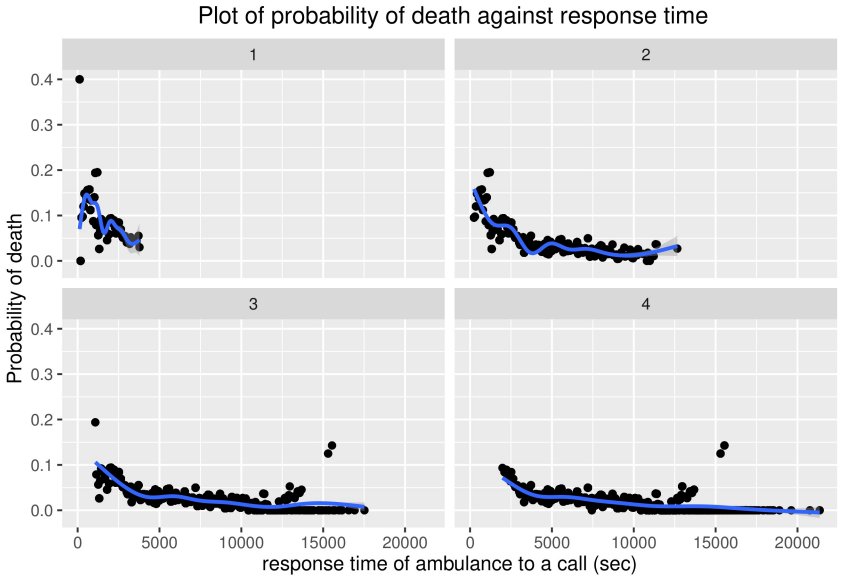
```
ambulance_hospital<- mutate(ambulance_hospital, "response time"=difftime(Arrival,Call))
Hospital_stay<-ambulance_hospital%>% group_by(`response time`) %>%
  summarise("mean length of stay"=mean(Length), "variance in length of stay"=var(Length))
ggplot(Hospital_stay,aes(x=`response time`, y= `mean length of stay`))+geom_line()+
  labs(title=" Does the response time and average length of stay have a correlation?",
       x="response time of ambulance to a call (sec)",
       y= "average length of stay")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(ambulance_hospital,aes(x=`response time`, y=Length)) + geom_point()+facet_wrap(~Category2)+
  labs(title=" Response time of each category",
       x="response time of ambulance to a call (sec)",
       y= "average length of stay")+
  theme(plot.title = element_text(hjust = 0.5))
```



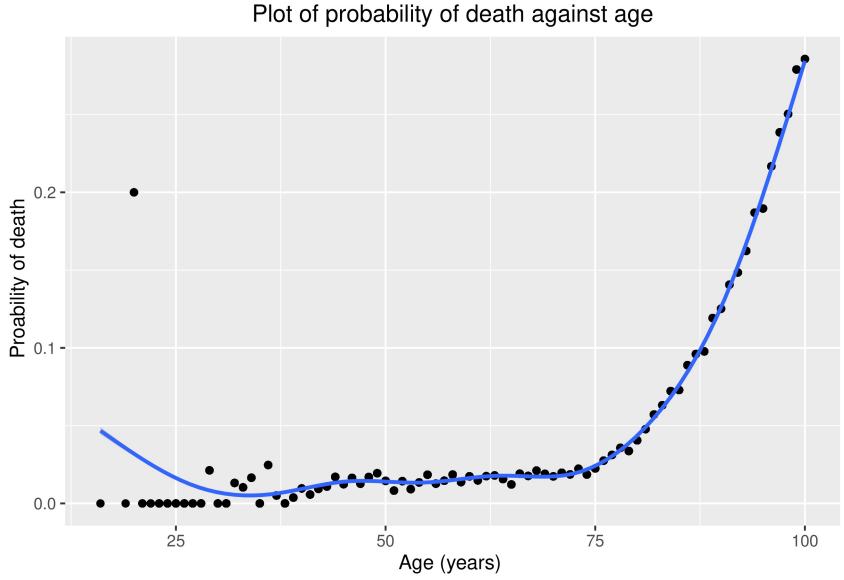

```
Death_rate<-ambulance_hospital%>% group_by(`response time`) %>%
  summarise("prob of death"=sum(Outcome==1)/(sum(Outcome==0) +sum(Outcome==1)),
            `response time`,Category2)
ggplot(Death_rate,aes(x=`response time`, y=`prob of death`)) + geom_point() + geom_smooth()+
  facet_wrap(~Category2)+
  labs(title=" Plot of probability of death against response time",
       x="response time of ambulance to a call (sec)",
       y= "Probability of death")+
  theme(plot.title = element_text(hjust = 0.5))
```



From these graphs we can see that there doesn't seem to be a correlation between the length of stay and the response time with the regression line being an average of 7-8, i also looked at the variance of the stay time and found that this was pretty constant until past the response time of 15000 seconds and before the response time of 2000 seconds, this is likely due to there being a much smaller sample size at this response

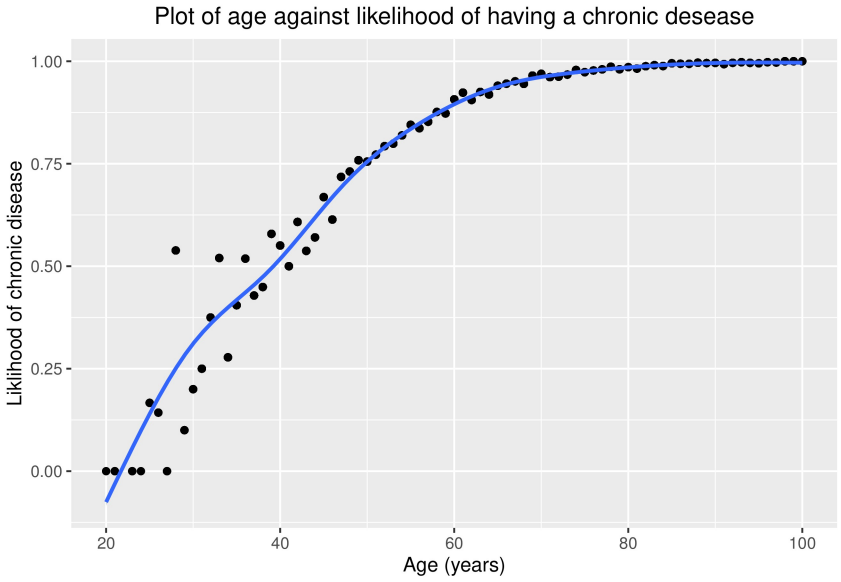
time. From the second plot this also confirms my suspicions as the response time increases for each category there still seems to be people staying for similar length of time. I find this strange as i would expect the shorter response time patients to stay for longer as they are more likely to be of a greater risk. However from the graph we can see that there is a correlation between response time and probability of death, this is that the faster the response time the more likely the outcome is death. this is probably due to the faster response times being for more critical patients. 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**] what are the dominant factor that contribute to age and the likelihood of death being so positively correlated?

```
Death_rate<-hospital%>% group_by(Age) %>%
  summarise("prob of death"=sum(Outcome==1)/(sum(Outcome==0) +sum(Outcome==1)),Age)
ggplot(Death_rate,aes(x=Age,y=`prob of death`))+geom_point()+geom_smooth()+
  labs(title="Plot of probability of death against age",
       x="Age (years) ",
       y= " Proability of death")+
  theme(plot.title = element_text(hjust = 0.5))
```

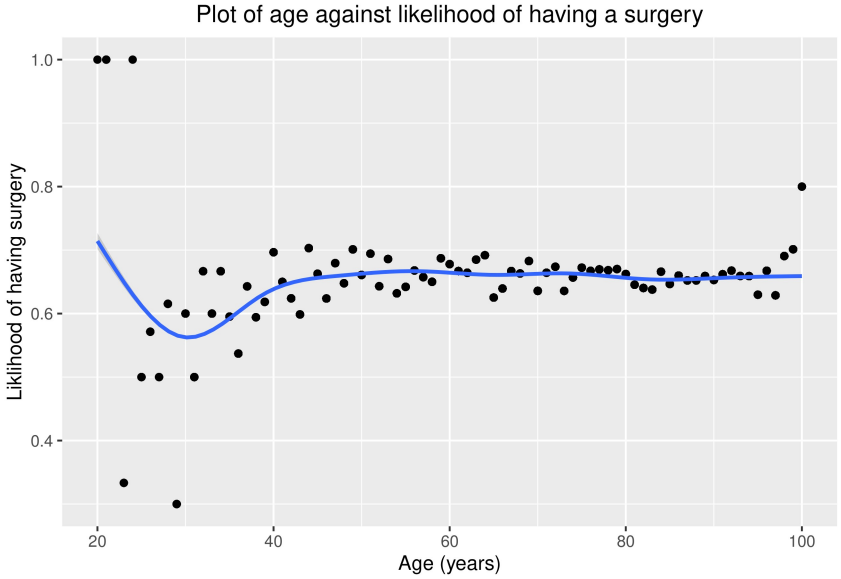


This graph indicates there is a clear positive correlation between the age of a patient and the death rate of the patient. As a patients age increases the likelihood of death in hospital also increases.

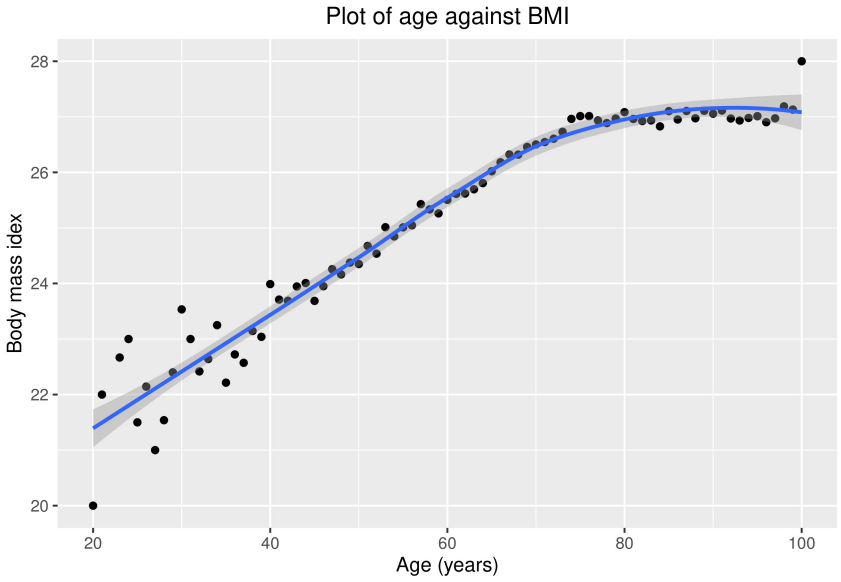
```
Chronic_likelihood<-ambulance_hospital%>% group_by(Age) %>%
  summarise("prob of chronic"=sum(Chronic==1)/(sum(Chronic==0) +sum(Chronic==1)),Length)
Surgery_likelihood<-ambulance_hospital%>% group_by(Age) %>%
  summarise("prob of surgery"=sum(Operation==1)/(sum(Operation==0) +sum(Operation==1)),Length)
BMI_mean<-ambulance_hospital%>% group_by(Age) %>%
  summarise("bmi mean"=mean(BMI))
ggplot(Chronic_likelihood,aes(x=Age,y=`prob of chronic`))+geom_point()+geom_smooth()+
  labs(title="Plot of age against likelihood of having a chronic desease",
       x="Age (years)",
       y= "Liklihood of chronic disease")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(Surgery_likelihood,aes(x=Age,y=`prob of surgery`))+geom_point()+geom_smooth()+
  labs(title="Plot of age against likelihood of having a surgery",
       x="Age (years)",
       y= "Liklihood of having surgery")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(BMI_mean,aes(x=Age,y=`bmi mean`))+geom_point()+geom_smooth()+
  labs(title="Plot of age against BMI",
       x="Age (years)",
       y= "Body mass idex")+
  theme(plot.title = element_text(hjust = 0.5))
```



From these three graphs you can see that the reason for the correlation between age and and death rate is dominated by the bmi and probability of having a chronic illness. A bmi past 25 is considered overweight and a bmi past 30 is considered obese however this is the average for people above the age of 50. There is a spike in probability of death of people who are aged around 23 this is likely and outlier. The probability of surgery seems to have no correlation with age indicating it doesn't effect the likelihood of death.

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 utopias health department from my finding i have become impressed with you ambulance and hospital service. From part C i can see that you have catagorised your emergencies and there is a prioritisation with your response times of your ambulances. however i would recommend finding out what type of accident has happened for example if the patient has been in a car crash, this is because this information can be used to alert other services such as the fire department or the police. currently there is no way of telling if you might need another service to get involved in the accident. It would make sense to create another column with which other service should also respond. From part A week can see the Frequency of call at each time of the day, this information can be used to identify when you will need more staff on rota.

From part B we know that as the length of stay and the probability of a patient being discharged are inversely proportionate however we don't know why this is the case, it might be useful to add a category suggesting if the patients condition worsened and how. lastly for part D i created graphs to find out the reason as to why age and death rate are so closely correlated however there are many more reason as to why this might be the case and it might be wise to take into account a patients previous injuries, there gender, if they are a smoker ect. All of these component contribute to a persons probability of dying and can be used to treat a patient more effectively. Lastly the problems that arose when using your data was that there were so many patients in the data base you would end up having multiple people having the same outcome this makes scatter plots unreliable therefore if you were to graph your data i would recommend using line plots more often.