MA20277 - Coursework 1

23720

```
library(dplyr)
library(lubridate)
library(ggplot2)
library(patchwork)
library(tidyr)
library(tibble)
```

Question 1

The Utopian charity *Respect for Pets* has collected data on cats and dogs for 1990-2023. Utopia only allows three dog breeds, Beagle, Dachshund and Maltese, and all pets have to be registered. The charity would like to gain some insight regarding the following questions:

The data for Question 1:

```
Pets <- read.csv("Pets.csv")
Cases <- read.csv("Cases.csv")
Maltese <- read.csv("Maltese.csv")
```

a) How has the number of dogs and cats changed over time? How has the popularity of the different dog breeds evolved since 1990?

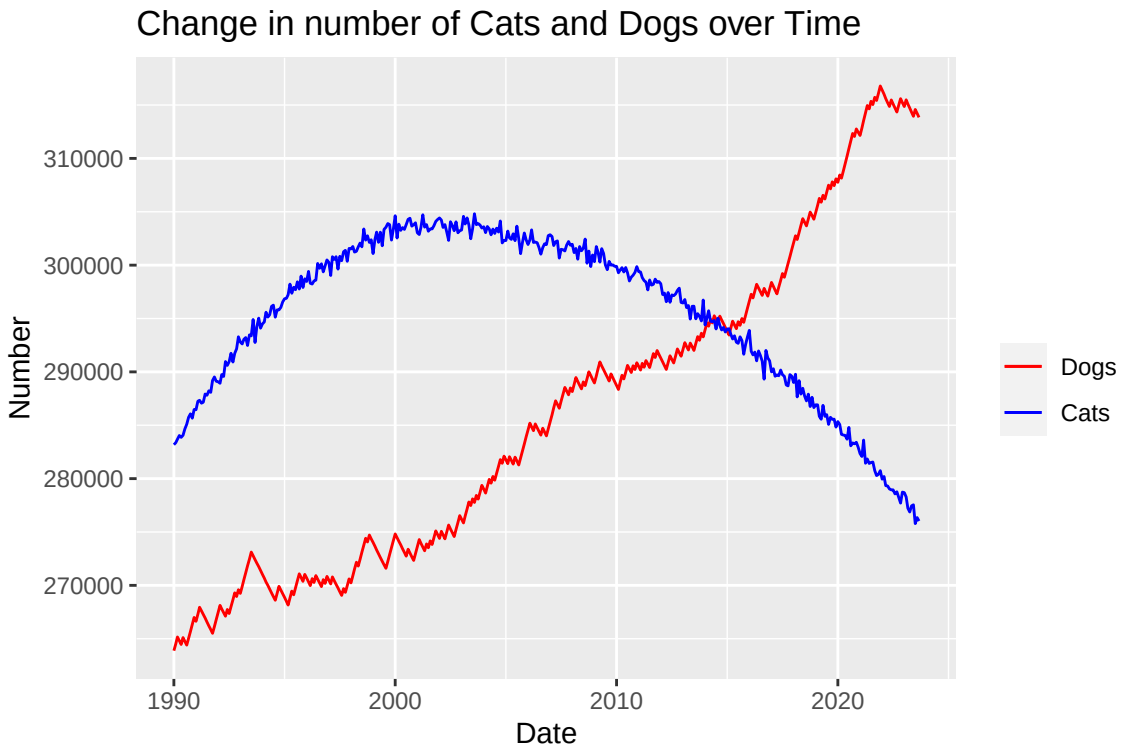
First I want to create one column for the date by combining the year and month columns and making it the correct format. Then I'm also going to create a new column to keep track of the total amount of dogs in Utopia.

```
Pets$Date <- paste(Pets$Year, Pets$Month, sep = "-")
Pets$Date <- as_date(Pets$Date , format = "%Y-%m")
```

```
Pets$TotalDogs <- Pets$Beagles+ Pets$Dachshund + Pets$Maltese
```

Now looking at the change in number of cats and dogs over time:

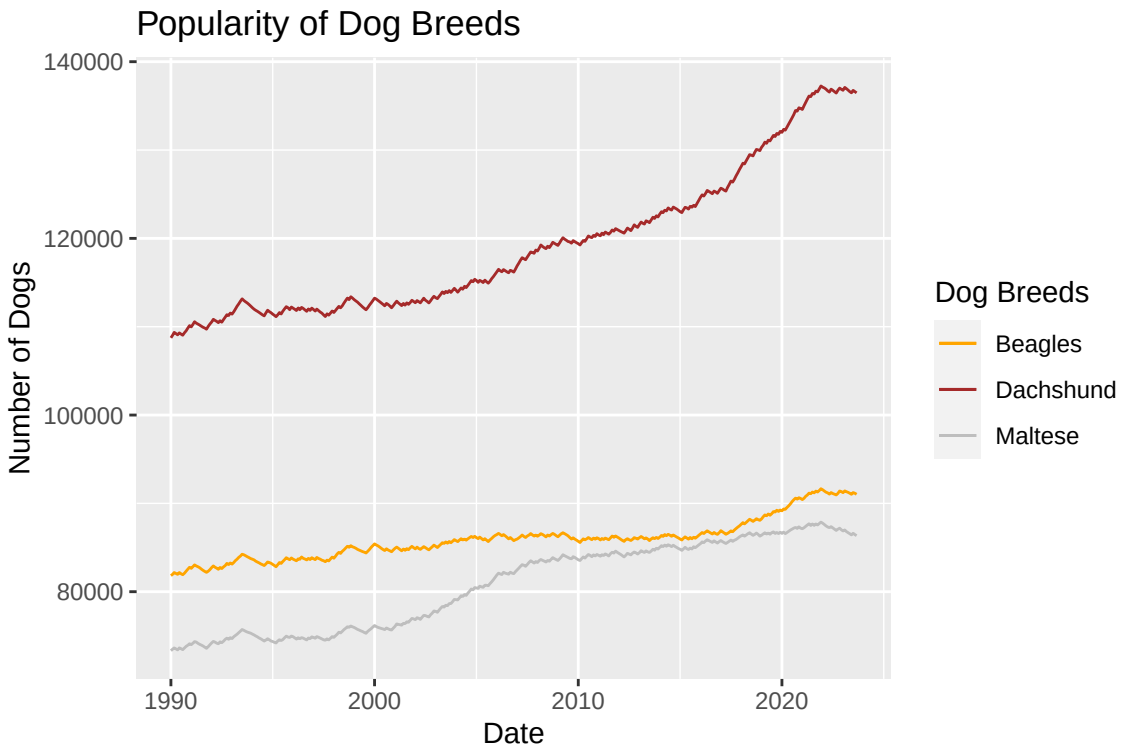
```
Pets%>%
  select("Date","TotalDogs", "Cats" )%>%
  ggplot(aes(x = Date)) + geom_line( aes(y =TotalDogs, colour = "Dogs"))+
  geom_line(aes(y =Cats, colour = "Cats")) +
  labs(title = "Change in number of Cats and Dogs over Time",
       y = "Number")+
  scale_colour_manual("", breaks = c("Dogs", "Cats"),
                      values = c("red", "blue"))
```



We see that the population of dogs has been increasing at a high rate. While numbers of cats increased until around 2004 where it peaked, and we've been seeing a downwards trend since, actually steeping below where the population was when the data was first collected.

Now looking at the popularity of dog breeds:

```
Pets%>%
  select("Date", "Beagles", "Dachshund", "Maltese")%>%
  ggplot(aes(x = Date)) + geom_line( aes(y= Beagles, colour = "Beagles")) +
    geom_line( aes(y = Dachshund, colour = "Dachshund")) +
    geom_line( aes(y = Maltese, colour = "Maltese")) +
    labs(title = "Popularity of Dog Breeds", y = "Number of Dogs")+
    scale_colour_manual("Dog Breeds", breaks = c("Beagles", "Dachshund", "Maltese"),
                        values = c("orange", "brown", "grey"))
```



We see that Dachshunds are, by quite a large margin, the most popular dog breed. Maltese had a bit of a surge in popularity in the 2000s, but never quite took over Beagles, who are the second most popular.

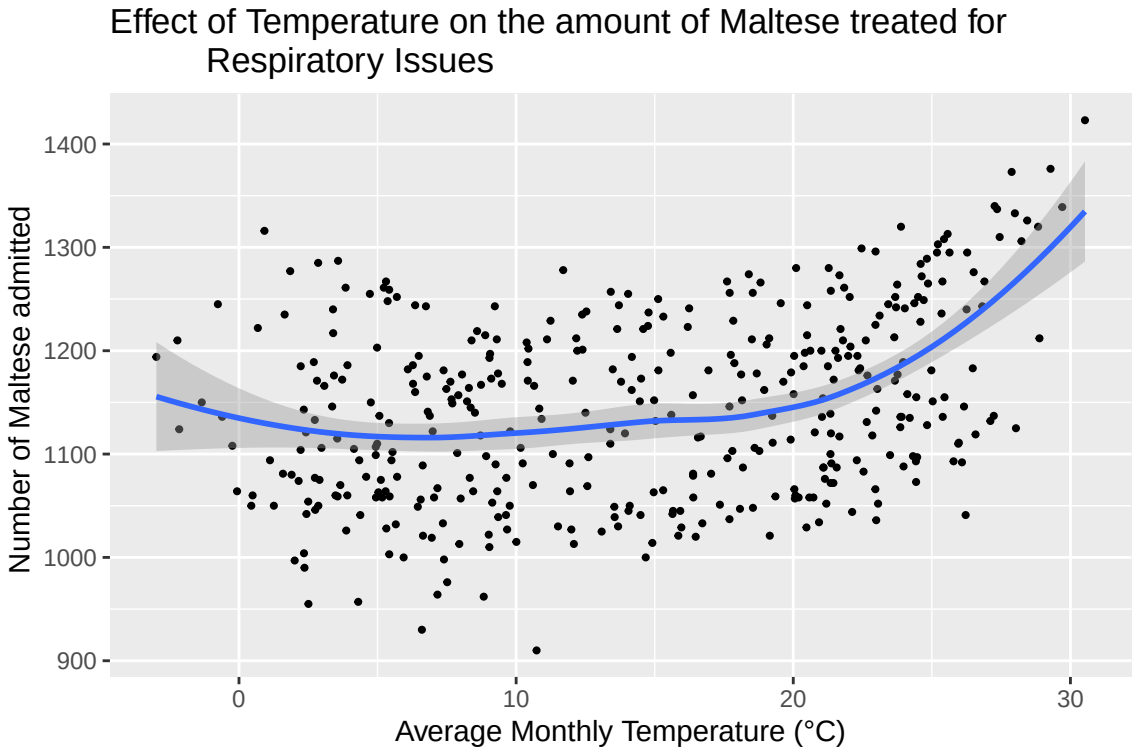
b) Maltese are known to experience respiratory issues, such as wheezing or asthma. How do environmental and physiological factors affect the risk of a Maltese experiencing these issues?

First lets look at Cases, this is the environmental impact on a Maltese's issues. There's some minor tweaking to do, we see the Date column is not in the correct format:

```
Cases$Date <- ymd(Cases$Date)
```

Now creating an average month temperature, then plotting against the number of incidences.

```
Cases%>%
  group_by(MonthYear =floor_date(Date, 'month'))%>%
  summarise(AvgMonTemp = mean(Temperature), TotalIssues = sum(Number))%>%
  ggplot(aes(x = AvgMonTemp, y =TotalIssues)) + geom_point(size = 0.75) + geom_smooth()+
    labs( title = "Effect of Temperature on the amount of Maltese treated for
          Respiratory Issues",
          x = "Average Monthly Temperature (°C)" ,
          y ="Number of Maltese admitted ")
```



I've converted the daily temperature into the Average Temperature for each month because we're working with temperatures over 30 years, it's clearer this way. We see in the more extreme temperatures, the more Maltese are treated for respiratory issues. Especially in high temperatures, when it starts going above 20°C. So temperature does impacts Maltese' respiratory issues. Which is what we expected: the higher the temperature, the hotter the dogs get which can cause wheezing and asthma.

Now lets look at Maltese data set which will tell us about physiological factors. I'm going to create a Weight-height Ratio which will combine the two variables.

```
Maltese$Ratio <- Maltese$Weight / Maltese$Height
Maltese%>%
  mutate(WeightHeight_Ratio = case_when(Ratio < 0.75 ~ "Too Low",
                            Ratio > 0.85 ~ "Too High",
                            Ratio>0.75 | Ratio<0.85 ~ "Good"))%>%
  group_by(WeightHeight_Ratio)%>%
  summarise(Percentage_with_Issues = 100*mean(RespiratoryIssues) )%>%
  arrange(Percentage_with_Issues)
```

```
## # A tibble: 3 x 2
##   WeightHeight_Ratio Percentage_with_Issues
##   <chr>                               <dbl>
## 1 Too Low                              12.1
## 2 Good                                 12.5
## 3 Too High                             13.4
```

I've classified each Maltese' health based on a ratio of their Weight and Height. Then I've calculated the percentage of each given health category that experienced respiratory issues. We see that the higher the ratio of weight to height, the higher the percentage of dogs with respiratory issues. This is what we expected, for example, dogs that are disproportionately heavy compared to their heights, are more likely to experience health problems. We also see that dogs that are more underweight-say are less likely to suffer with these respiratory issues, but they might suffer with issues such as tiredness or weakness.

Question 2

The Utopian Fire Department has gathered data on their activities for 2022. They also managed to provide you with access to some data for the houses in Utopia. The Utopian Fire Department asks you to address the following questions:

The data for Question 2:

```
Fires<- read.csv("Fires.csv")
HousingReg <- read.csv("Housing Register.csv")
```

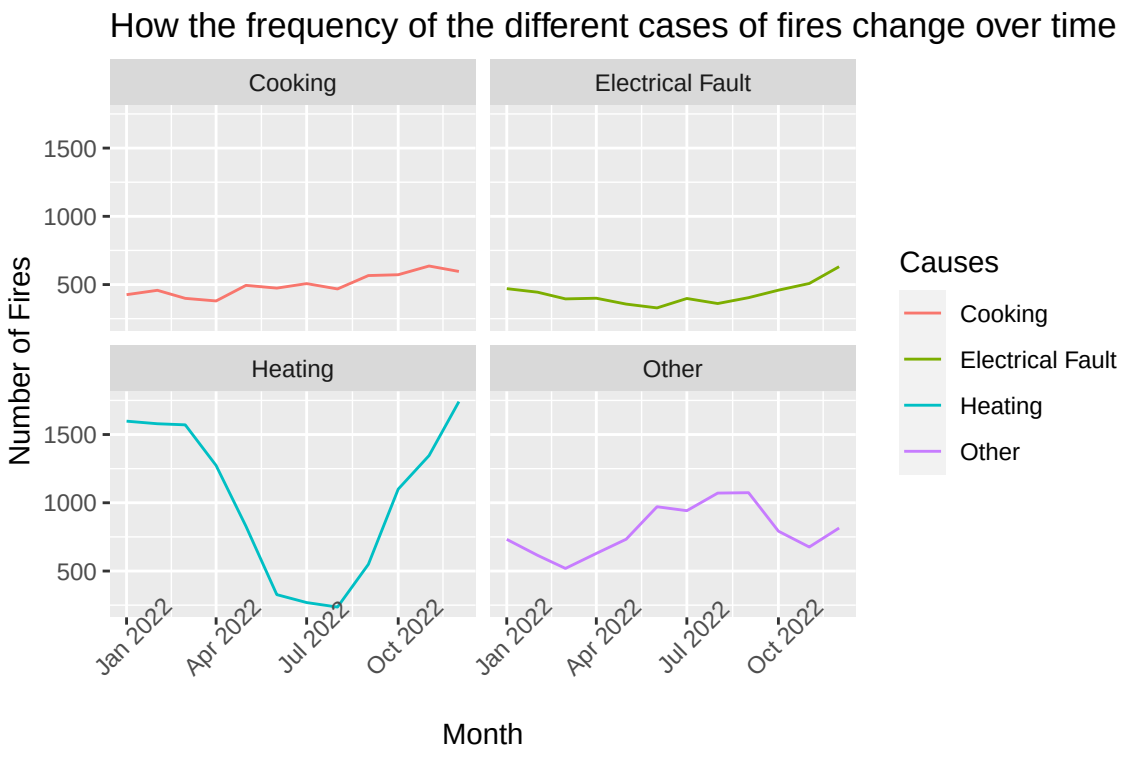
a) How does the frequency of the different causes for fires vary other time? How many casualties were attributed to each cause and are there differences in the frequency with which casualties occur across causes?

First we see that the Date variable isn't the right type so:

```
Fires$Date <- as_datetime(Fires$Date, format = "%Y-%m-%d %H:%M")
```

Looking at the frequency of causes for fires over time, but counting the total fires per month:

```
Fires%>%
  group_by(MonthYear =floor_date(Date, 'month'))%>%
  select("MonthYear","Cause" )%>%
  count(Cause)%>%
  ggplot(aes(x = MonthYear, y = n , colour = factor(Cause))) + facet_wrap(~Cause)+geom_line() +
  labs( title = "How the frequency of the different cases of fires change over time",
        y = "Number of Fires", x = "Month", colour ="Causes") +
  theme(axis.text.x = element_text(angle = 45))
```



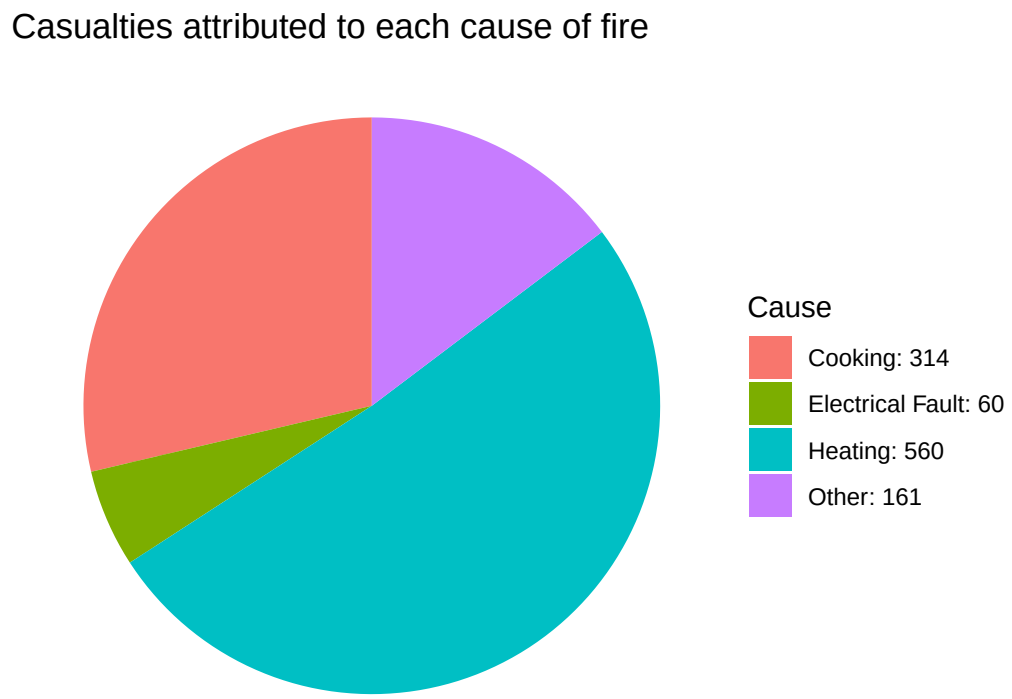
This shows us that Cooking fires stay quite constant over the year, we expect this as the amount of cooking doesn't really change with seasonality. Where we see the most obvious difference is Heating, there are a lot less fires over the summer, due to the heating not being on in almost all homes. We then see an increase when

temperatures drop in the winter. This is also seen slightly in Electrical faults fires, but not to the extent of heating, but this makes sense again, we expect less electrics (like lighting) on during the summer. For the fires classified as Other, we see an increase over the summer months, this could be explained by barbecues for example.

Now how many casualties were linked to each cause of fire:

```
NumofCas <- Fires%>%
  group_by(Cause)%>%
    filter(Casualties!= 0)%>%
  summarise(Casualties = n())

ggplot(NumofCas, aes( x = "", y = Casualties, fill = Cause)) +   geom_col() +
  coord_polar(theta = "y") +
  labs(x = "", y = "", title = "Casualties attributed to each cause of fire")+
  scale_fill_discrete(labels = c("Cooking: 314", "Electrical Fault: 60",
                                 "Heating: 560", "Other: 161")) +
  theme_void()
```



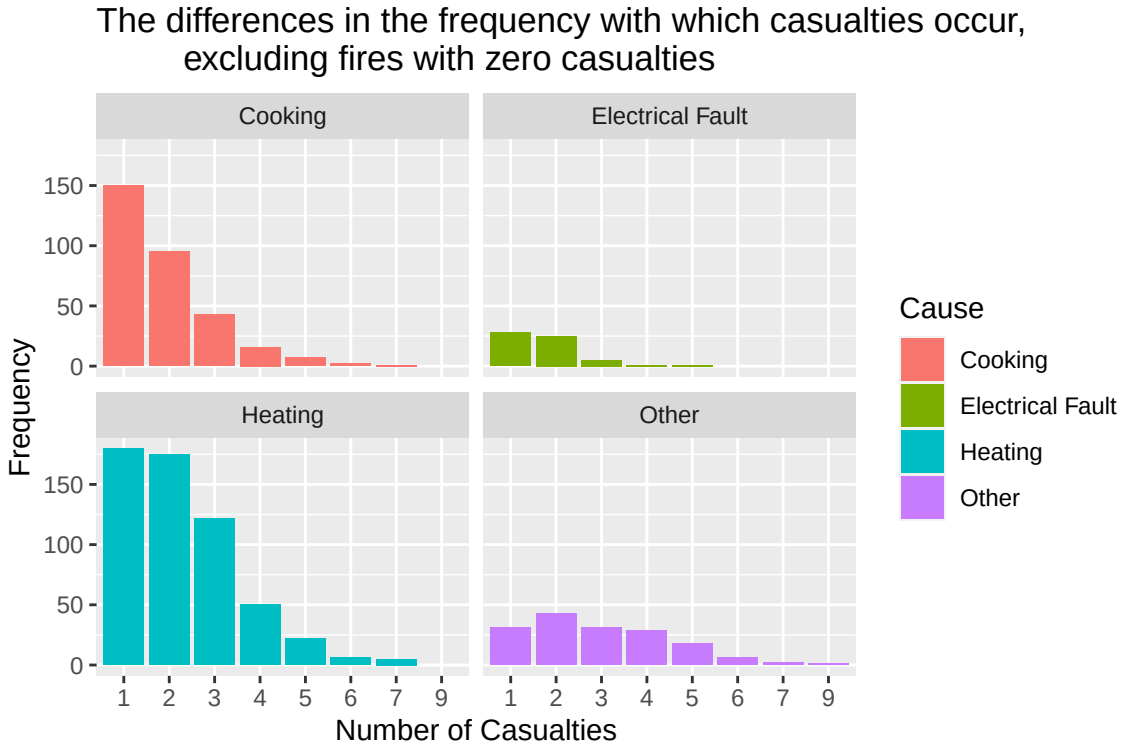
We see that Heating fires has the largest amount of casualties, taking up over 50% of all the deaths. The second most frequent is Cooking fires, while Electrical faults have the fewest.

But are there differences in the frequency with which casualties occur across causes?

In the majority of fires, there are no casualties, but this makes it difficult to look at when there are casualties. For this reason, in this question, I'm only going to look at fires where we do have casualties, so we can see the spread better.

```
Fires%>%
  filter(Casualties!= 0)%>%
  group_by(Casualties, Cause)%>%
  count(Casualties)%>%
```

```
ggplot(aes(x = as.factor(Casualties), y = n, fill = Cause))+
  facet_wrap(~Cause) +geom_bar(stat = "identity") +
  labs(title = "The differences in the frequency with which casualties occur,
       excluding fires with zero casualties",
       x = "Number of Casualties",
       y = "Frequency")
```



Here we see what the distribution of casualties for each different cause looks like when we only look at fires that result in casualties. We see that with cooking and heating fires, most of the casualties fall between 1 and 3. While with Electrical Fault fires, its very unlikely to have casualties greater than 2. With those in the "Other" category, there are a few more fires with larger casualties.

But we can also look at the average mortality rate:

```
Fires%>%
  filter( Casualties!=0)%>%
  group_by(Cause)%>%
  summarise(AvgCasualties = mean(Casualties))%>%
  arrange(desc(AvgCasualties))
```

```
## # A tibble: 4 x 2
##   Cause            AvgCasualties
##   <chr>                    <dbl>
## 1 Other                     2.95
## 2 Heating                   2.28
## 3 Cooking                   1.87
## 4 Electrical Fault          1.7
```

Surprisingly, we see that when there are casualties, "Other" fires have the highest average mortality rate, which contradicts our pie chart from earlier, but this is just due to it's wider breadth of casualties. It is unlike Heating and Cooking fires, where it is very frequent to have 1-3 casualties. As mentioned, the larger

casualty occurances pull the average up.

b) Are there any differences in the risk of fire for the different types of property? What is the relation between the year a property was built and the risk of fire?

Joining the data sets:

```
Fires <- rename(Fires, "ID" = RegisterNumber)
Fires_joined <- inner_join(Fires, HousingReg, by = "ID")
```

I'm going to define a measure of risk of fire as each properties possession of smoke alarms and carbon monoxide alarms, as without these a fire could go undetected by home-owners and hence become a problem where the fire brigade has to be called. So the higher the "Average Safety Score" the less risk of a fire.

```
Fires_joined$SafeScore = Fires_joined$Smoke + Fires_joined$CO
```

Looking at risk across different types of properties:

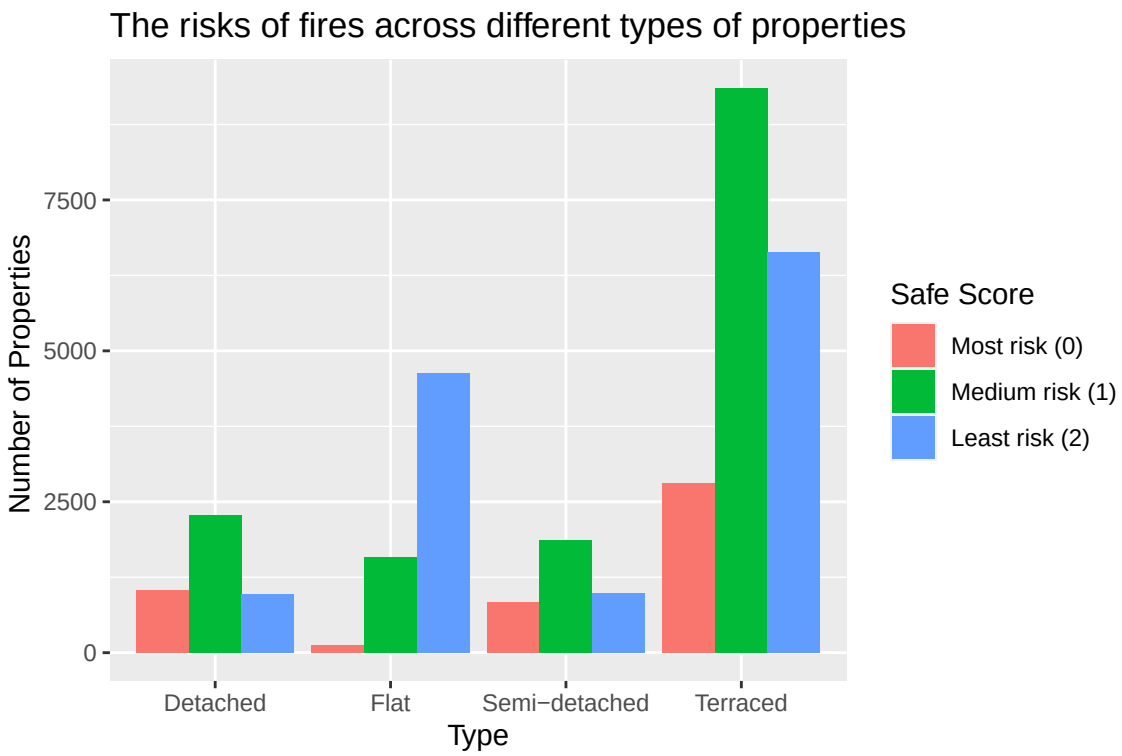
```
Fires_joined%>%
  select("Date", "SafeScore", "Type")%>%
  group_by(Type)%>%
  summarise(AvgSafeScore = mean(SafeScore))%>%
  arrange(AvgSafeScore)
```

```
## # A tibble: 4 x 2
##   Type          AvgSafeScore
##   <chr>                <dbl>
## 1 Detached             0.983
## 2 Semi-detached        1.04
## 3 Terraced             1.20
## 4 Flat                 1.71
```

Here we see that Flats have the lowest risk of fire, while Detached houses have the highest risk of fire.

Here I display the differences in graphical format:

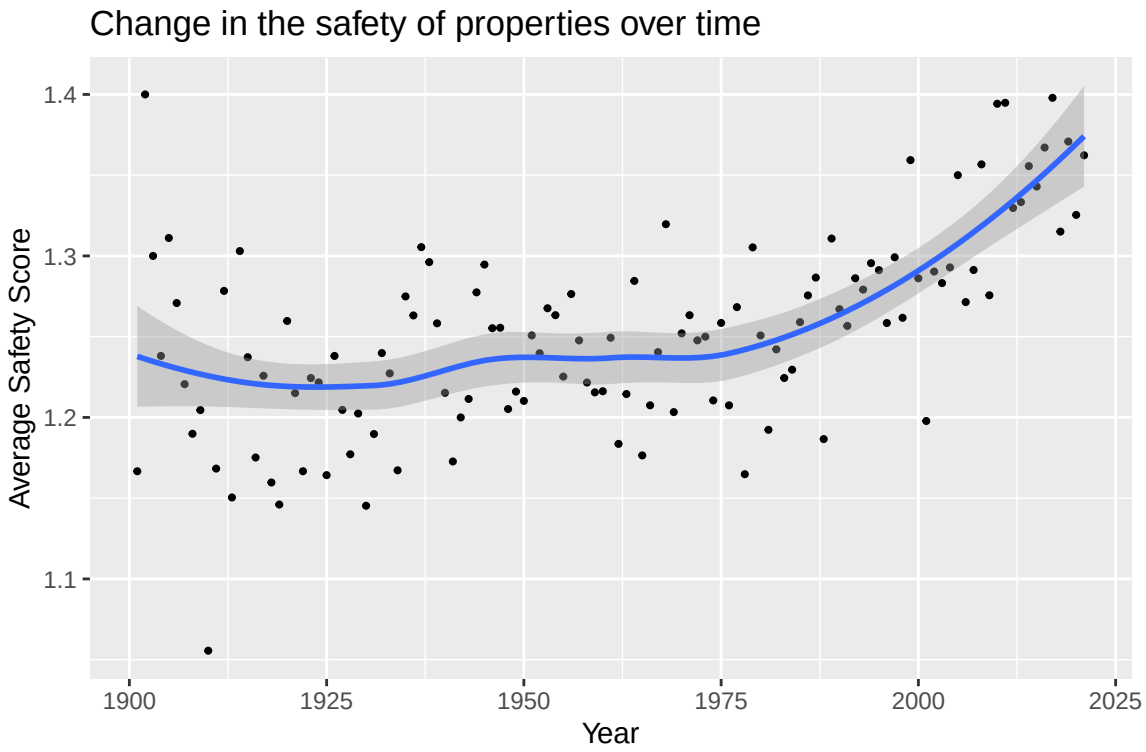
```
Fires_joined%>%
  group_by(Type, SafeScore)%>%
  summarise(Number = n())%>%
  ggplot(aes(fill = as.character(SafeScore),
             x = Type, y = Number)) +
    geom_bar(position = "dodge", stat = "identity") +
    labs(title = "The risks of fires across different types of properties",
         fill = "Safe Score", y = "Number of Properties ")+
  scale_fill_discrete(labels = c("Most risk (0)", "Medium risk (1)", "Least risk (2)"))
```



This graph empahsises the really low risk that flats have, while the bars for Detached and Terraced can be quite deceiving. It appears that Terraced has a larger amount of properties with a low "Safe Score" but that's just because there are more Terraced houses. Therefore its better to look at the averages.

Looking at risk based on age of the property:

```
Fires_joined%>%
  select("Year", "SafeScore")%>%
  group_by(Year)%>%
  summarise(AvgSafeScore = mean(SafeScore))%>%
  ggplot(aes(x = Year, y  = AvgSafeScore)) + geom_point(size = 0.75) + geom_smooth() +
    labs( title = "Change in the safety of properties over time", y = "Average Safety Score")
```

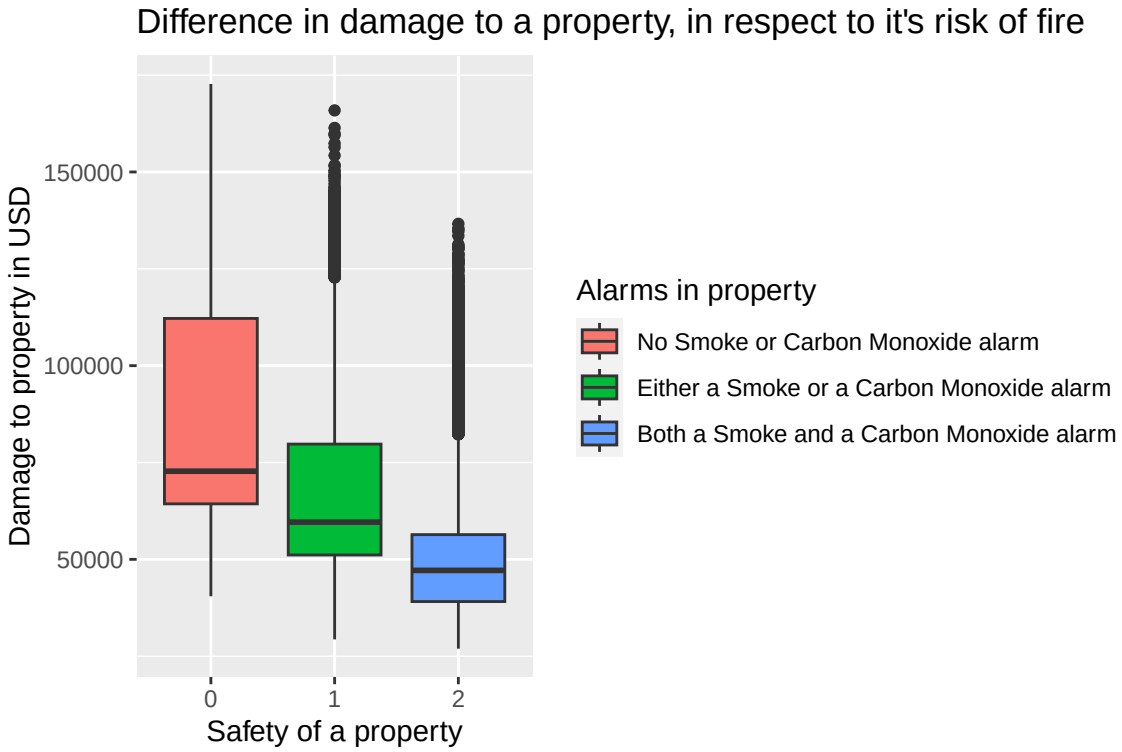


We see that over time, safety has increased, so risk of fire has decreased. This is probably due to more laws around new-builds and fire safety. Overall, newer houses are at less risk of fire.

c) The Fire Department wishes to run a campaign which encourages home owners to install smoke and carbon monoxide detectors. What does the data reveal regarding the benefits of installing smoke and carbon monoxide detectors?

Again we'll use our safescore variable. But compare it to damages, and make it a bit more accessible to the public, this way the plot can be displayed on posters etc:

```
Fires_joined%>%
  select("SafeScore", "Damage")%>%
  ggplot(aes(x = as.factor(SafeScore), y = Damage, fill = as.factor(SafeScore))) +
  geom_boxplot() +
  labs(title = "Difference in damage to a property, in respect to it's risk of fire",
       x = "Safety of a property",
       y = "Damage to property in USD",
       fill = "Alarms in property") +
  scale_fill_discrete(labels = c("No Smoke or Carbon Monoxide alarm",
                                 "Either a Smoke or a Carbon Monoxide alarm",
                                 "Both a Smoke and a Carbon Monoxide alarm"))
```



This makes the benefits of smoke and carbon monoxide detectors really clear. It shows that no presence predicts a lot more damages to the property. Adding one detector will decrease the damages, while having both a smoke and a carbon monoxide alarm is the best thing you can do to your property to reduce the risk of a fire, and it's damages.