

MA20277 - Coursework 1

Harry Rogers

```
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:

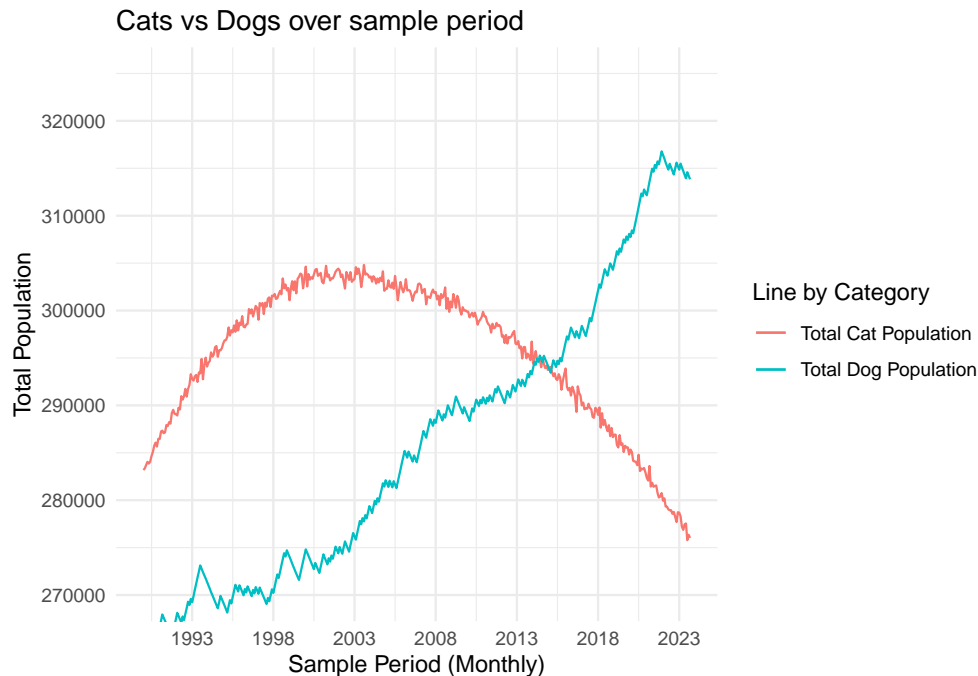
- a) How has the number of dogs and cats changed over time? How has the popularity of the different dog breeds evolved since 1990?

```
pets = read.csv("data/Pets.csv")
maltese_Dogs = read.csv("data/Maltese.csv")
respiratory_Cases = read.csv("data/Cases.csv")
```

1A - Cats and Dogs Population Analysis

```
Cats_Dogs_Aggregated_Data = pets %>% mutate(Total_Dogs = (Beagles + Dachshund + Maltese),
                                             Date_col = make_date(Year,Month,1)) %>%
# Add columns calculating total for Cats and Dogs,
# and combining year and month columns using lubridate
select(Date_col,Cats,Total_Dogs)
# Selecting only needed columns for analysis

ggplot(Cats_Dogs_Aggregated_Data,aes(Date_col)) +
  geom_line(aes(y = Cats, color = "Total Cat Population")) +
  geom_line(aes(y = Total_Dogs, color = "Total Dog Population")) +
  theme(legend.title = element_text(face = "bold")) +
  labs(colour = "Line by Category", title = "Cats vs Dogs over sample period") +
  xlab("Sample Period (Monthly)") +
  ylab("Total Population") +
  coord_cartesian(xlim = c(make_date(1990,1,1),make_date(2023,9,1)),
                  ylim = c(270000,325000)) + theme_minimal() +
  scale_x_date(date_breaks = "5 years",date_labels = "%Y")
```

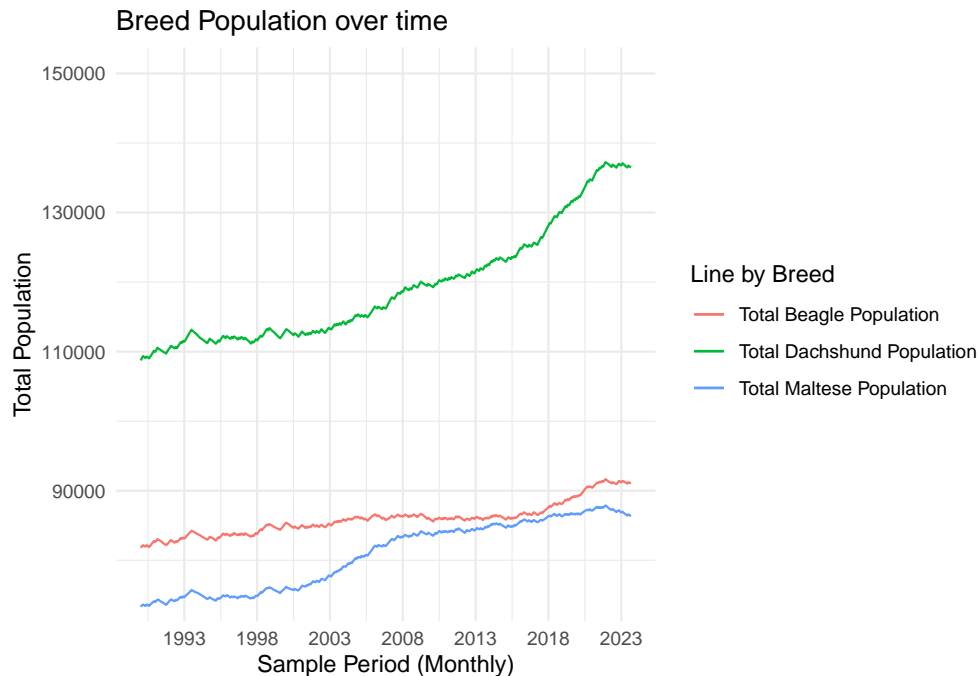


We are able to see that while cats used to have a higher population than dogs within our data in the period 1990 to 2014, however dogs have a much higher population than cats post 2014. The population of cats has been steadily decreasing since the early 2000's.

1A - Breed population analysis

```
Breed_Analysis = pets %>% mutate(Date_col = make_date(Year,Month,1)) %>%
  select(Date_col,Beagles,Dachshund,Maltese)

ggplot(Breed_Analysis,aes(Date_col)) +
  geom_line(aes(y = Beagles, color = "Total Beagle Population")) +
  geom_line(aes(y = Dachshund, color = "Total Dachshund Population")) +
  geom_line(aes(y = Maltese, color = "Total Maltese Population")) +
  ggtitle("Breed Population over time") +
  labs(colour = "Line by Breed",
       x = "Sample Period (Monthly)",
       y = "Total Population") +
  coord_cartesian(xlim = c(make_date(1990,1,1),make_date(2023,9,1)),
                  ylim = c(75000,150000)) + theme_minimal() +
  scale_x_date(date_breaks = "5 years",date_labels = "%Y")
```



We are able to see that Maltese and Beagle breeds have stayed consistently similar throughout the sample period, with a mean of 82,000 and 87,000 respectively. The Dachshund breed has been higher throughout the sample period and continues to grow, reaching a maximum of 141,000 and a mean of 123,000. If more analysis was required, we could also consider looking at the proportion of rescue dogs and cats compared to their population. This would provide information on their popularity during ownership.

- 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?

By conducting preliminary analysis of the data, we are able to determine a periodic cycle in cases and temperature over each year. We can therefore analyse cases and temperature by grouping the data by month, and analyse the mean temperature and cases monthly.

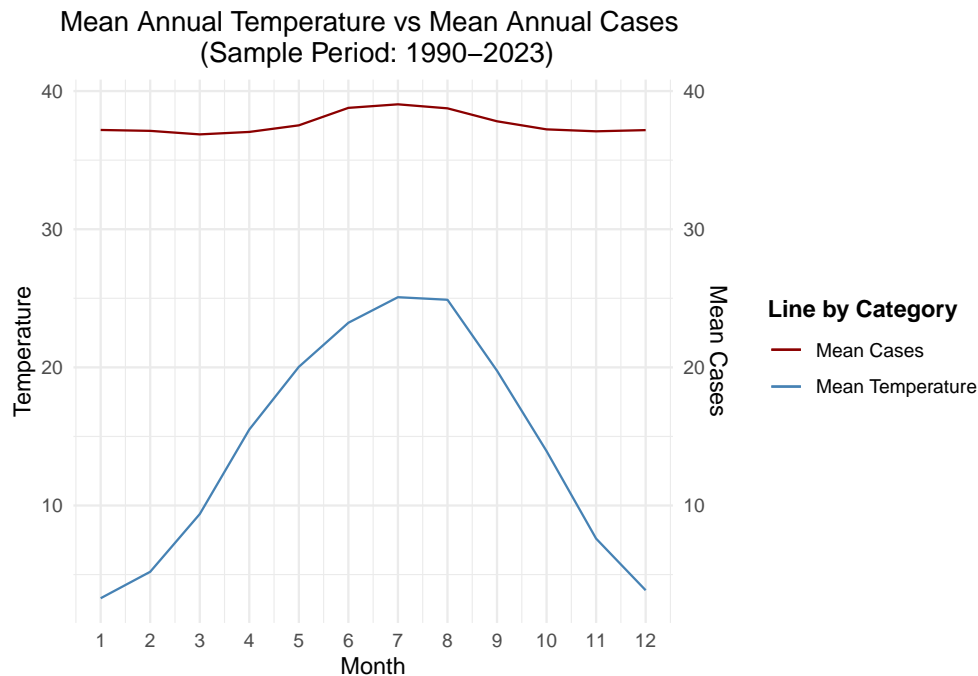
1B - Maltese Respiratory Issue Analysis

```
respiratory_Cases_Clean = respiratory_Cases %>%
mutate(Date = ymd(Date))

respiratory_Cases_Monthly = respiratory_Cases_Clean %>%
  mutate(Date = month(Date)) %>%
  group_by(Date) %>%
  summarise('Mean_Temp' = mean(Temperature),
            'Mean_Cases' = mean(Number))

ggplot(respiratory_Cases_Monthly, aes(x = Date)) +
  geom_line(aes(y = Mean_Temp, color = "Mean Temperature")) +
  geom_line(aes(y = Mean_Cases, color = "Mean Cases")) +
  scale_color_manual(values = c("Mean Temperature" = "steelblue",
                                "Mean Cases" = "darkred")) +
```

```
labs(x = "Month",
     y = "Temperature",
     colour = "Line by Category") +
theme_minimal() +
theme(plot.title = element_text(hjust = 0.5),
      legend.title = element_text(face = "bold")) +
ggtitle("Mean Annual Temperature vs Mean Annual Cases \n (Sample Period: 1990-2023)") +
scale_y_continuous(sec.axis = sec_axis(~ ., name = "Mean Cases")) +
scale_x_continuous(limits = c(1, 12), breaks = seq(1, 12, by = 1))
```



```
# Correlation between temperature and number of cases throughout the entire sample period
cor(respiratory_Cases$Temperature, respiratory_Cases$Number)
```

```
## [1] 0.1425982
```

We use monthly mean temperature vs monthly mean cases not to analyse a trend between temperature and respiratory cases, but instead to analyse the seasonality of both occurring.

We may conclude that while cases increase when temperature increases, this may not be the causation, and could be down to something else such as allergies or increased respiration rates at that time of year. This is reinforced by the correlation coefficient between temperature and cases being around 0.14, suggesting no clear correlation.

1B - Box plots

```
maltese_Dogs = mutate(maltese_Dogs, Respiratory_Issues = case_when(
  RespiratoryIssues == 1 ~ "Respiratory Issues",
  RespiratoryIssues == 0 ~ "No Respiratory Issues"))
```

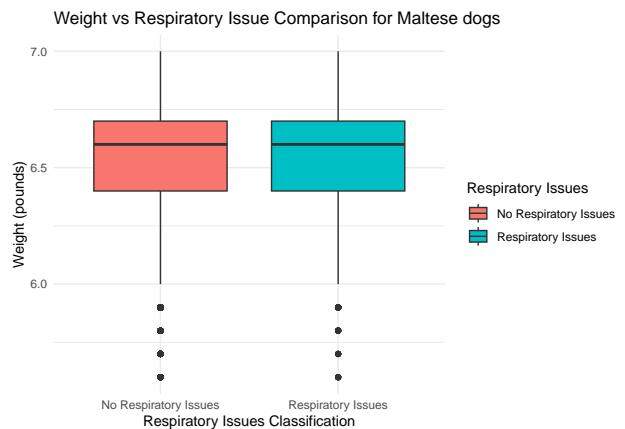
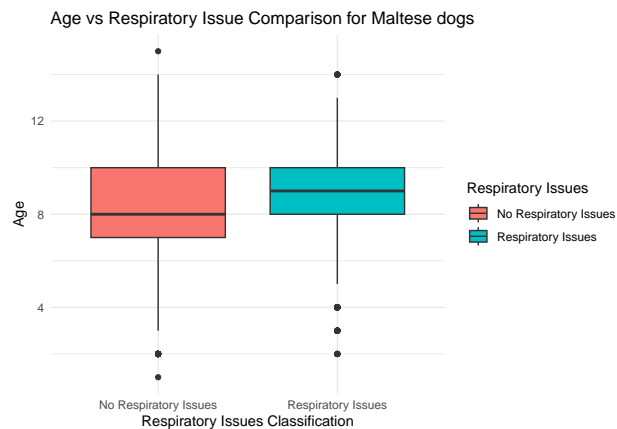
```

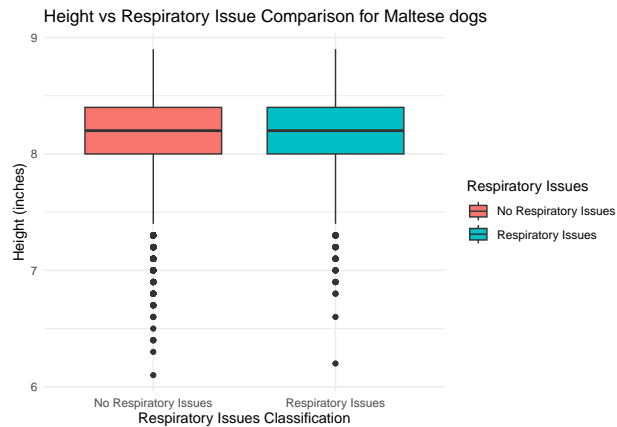
ggplot(maltese_Dogs,
      aes(x = Respiratory_Issues, y = Age, fill = Respiratory_Issues)) +
geom_boxplot() +
theme_minimal() +
labs(title = "Age vs Respiratory Issue Comparison for Maltese dogs",
      x = "Respiratory Issues Classification",
      y = "Age",
      fill = "Respiratory Issues")

ggplot(maltese_Dogs,
      aes(x = Respiratory_Issues, y = Weight, fill = Respiratory_Issues)) +
geom_boxplot() +
theme_minimal() +
labs(title = "Weight vs Respiratory Issue Comparison for Maltese dogs",
      x = "Respiratory Issues Classification",
      y = "Weight (pounds)",
      fill = "Respiratory Issues")

ggplot(maltese_Dogs,
      aes(x = Respiratory_Issues, y = Height, fill = Respiratory_Issues)) +
geom_boxplot() +
theme_minimal() +
labs(title = "Height vs Respiratory Issue Comparison for Maltese dogs",
      x = "Respiratory Issues Classification",
      y = "Height (inches)",
      fill = "Respiratory Issues")

```





While we can see that analysing height and weight for both classifications shows no significant difference, when we look at age, we are able to draw a conclusion that generally higher aged dogs are more susceptible to respiratory issues, with a higher median, Q3 and smaller IQR. We are however limited by our data, since we only have 2,500 observations with Respiratory issues, compared to 17,000 without any.

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:

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

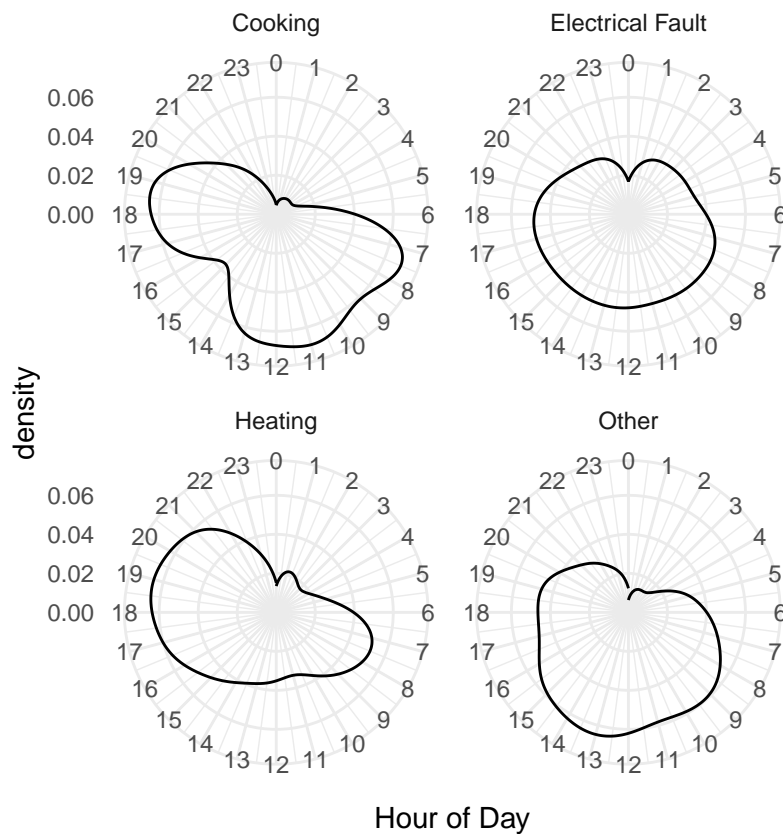
```
Fires = read.csv("data/Fires.csv")
Housing_Reg = read.csv("data/Housing Register.csv")
```

2A - Hourly density plots

```
Fires = Fires %>% mutate(Date_Clean = ymd_hm(Date),
  minute = minute(Date_Clean),
  hour = hour(Date_Clean),
  day = day(Date_Clean),
  month = month(Date_Clean),
  year = year(Date_Clean),
  time = hour + (minute/60),
  Casualties_Involved = case_when(
    Casualties > 0 ~ "Casualties Involved",
    Casualties == 0 ~ "No Casualties Involved"))

ggplot(Fires, aes(x = time)) + coord_polar(theta = 'x') +
  geom_density() +
  labs(x = 'Hour of Day',
  Title = "A plot showing the frequency of fires each day") +
  facet_wrap(vars(Cause)) +
```

```
scale_x_continuous(breaks = seq(0,23,1),limits = c(0,24)) +  
theme_minimal()
```



From the density function graphs above, we are able to analyse that each category has a different probability density function over a given 24 hrs. Cooking has the highest frequencies between the ranges 7am-8am, 11am-1pm and 5pm to 8pm; all times associated to breakfast, lunch and dinner. Electrical faults are evenly distributed over 8am to 7pm, Heating has the highest density over 5pm to 9pm, and Other has no identifiable trends apart from being evenly distributed during sociable hours.

2A - Casualties by causation

```
# Summarise the data by Cause, and calculate some fields for analysis.  
Casualties_by_causation_analysis = Fires %>%  
  group_by(Cause) %>%  
  summarise(Total_Casualties = sum(Casualties),  
            Count_Casualties = sum(Casualties > 0),  
            Total_Incidents = n(),  
            risk_likelihood = 100*(Count_Casualties/Total_Incidents))
```

Note, care must be taken to distinguish the difference between there being a casualty (a count) and the sum of casualties. Both are useful for analysis, however for the proportion of incidents where casualties have been involved, we will use the count.

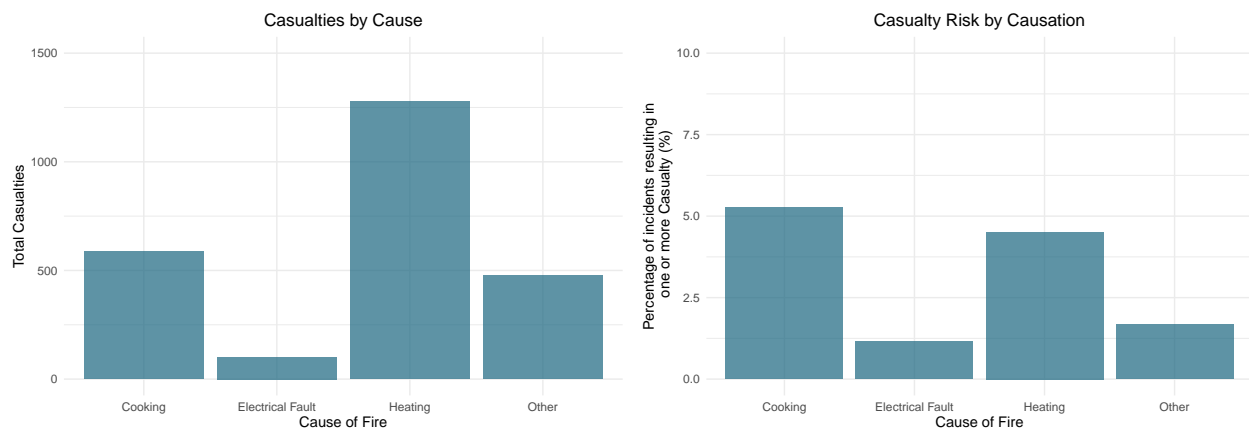
```

# Total Casualties by cause of fire
ggplot(Casualties_by_causation_analysis, aes(x = Cause, y = Total_Casualties)) +
  geom_bar(stat = "identity", fill=rgb(0.1,0.4,0.5,0.7)) +
  theme_minimal() +
  labs(x = "Cause of Fire",
       y = "Total Casualties",
       title = "Casualties by Cause") +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_cartesian(ylim = c(0,1500))

# Showing the proportion of casualties vs incidents

ggplot(Casualties_by_causation_analysis, aes(x = Cause, y = risk_likelihood)) +
  geom_bar(stat = "identity", fill=rgb(0.1,0.4,0.5,0.7)) +
  theme_minimal() +
  labs(x = "Cause of Fire",
       y = "Percentage of incidents resulting in\n one or more Casualty (%)",
       title = "Casualty Risk by Causation") +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_cartesian(ylim = c(0,10))

```

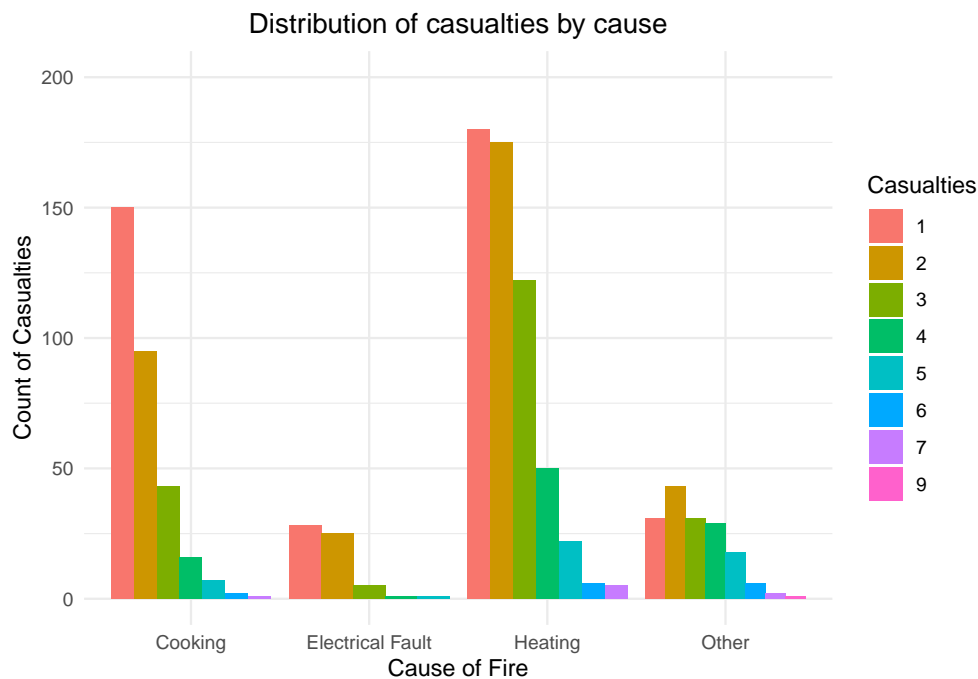


By examining total casualties by cause, and their likelihood to involve casualties, we can examine while Heating fires tend to have the most casualties involved, cooking fires are more likely to involve casualties.

2A - Casualty Distribution

```
# Calculate Casualty distribution, filtering on where casualties occur.
Casualty_Distribution = Fires %>% filter(Casualties > 0) %>%
  group_by(Cause,Casualties) %>%
  summarise(Count_Of_Incidents = n())

# Distribution of casualties
ggplot(Casualty_Distribution,
       aes(fill = as.character(Casualties),
           x = Cause,
           y = Count_Of_Incidents)) +
  geom_bar(position="dodge",stat = "identity") +
  theme_minimal() +
  labs(x = "Cause of Fire",
       y = "Count of Casualties",
       title = "Distribution of casualties by cause") +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_cartesian(ylim = c(0,200)) +
  scale_fill_discrete(name = "Casualties")
```



This is further examined in the distribution of casualties by cause, where heating fires tend to have a higher number of casualties involved.

- 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?

```
# Join housing registry to fire dataset.
Fire_join = Fires %>% left_join(Housing_Reg, by = c("RegisterNumber" = "ID"))
```

2B - Building Analysis

```
# Summarise data for building analysis

Building_Count = Housing_Reg %>% group_by(Type) %>%
  summarise(Count_Of_Properties = n())

knitr::kable(Building_Count, caption = "A count of buildings in Housing_Reg")
```

Table 1: A count of buildings in Housing_Reg

Type	Count_Of_Properties
Detached	544901
Flat	701857
Semi-detached	466531
Terraced	1759683

```
Building_Type_Analysis = Fire_join %>% group_by(Type) %>%
  summarise(Total_Fires = n(),
    Fires_Involving_Casualties = sum(Casualties > 0),
    Fires_Involving_Casualties_Perc = 100*Fires_Involving_Casualties/Total_Fires) %>%
  left_join(Building_Count, by = c("Type"="Type")) %>%
  mutate(Fires_Per_Building_Perc = 100*Total_Fires/Count_Of_Properties)
```

```
# Analyse Total fires by building type
ggplot(Building_Type_Analysis, aes(x = Type, y = Total_Fires)) +
  geom_bar(stat = "identity", fill=rgb(0.4,0,0.48627,0.7)) +
  theme_minimal() +
  labs(x = "Building Type",
    y = "Total Fires",
    title = "Fires by Building Type") +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_cartesian(ylim = c(0,20000))
```

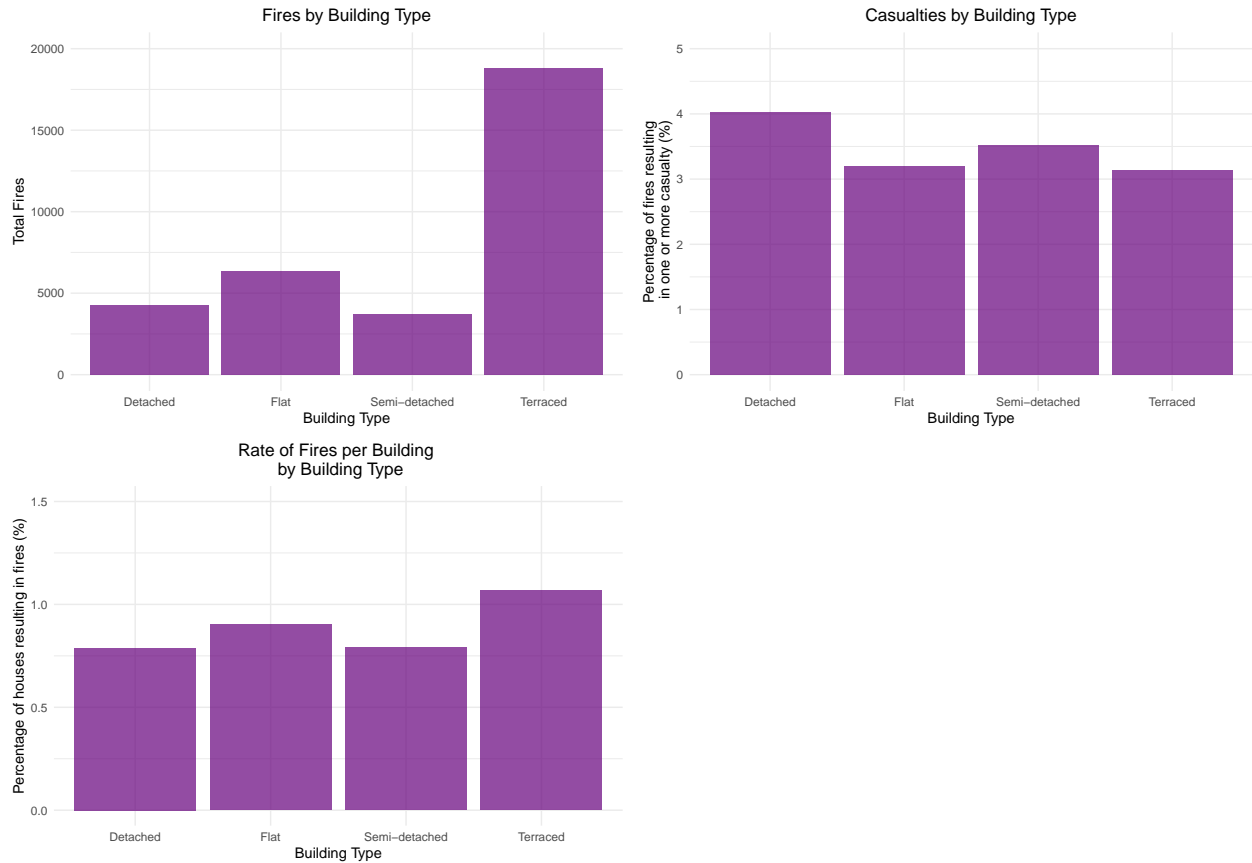
```
# Analyse casualties by building type
ggplot(Building_Type_Analysis, aes(x = Type, y = Fires_Involving_Casualties_Perc)) +
  geom_bar(stat = "identity", fill=rgb(0.4,0,0.48627,0.7)) +
  theme_minimal() +
  labs(x = "Building Type",
    y = "Percentage of fires resulting\n in one or more casualty (%)",
    title = "Casualties by Building Type") +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_cartesian(ylim = c(0,5))
```

```
# Percentage of houses resulting in fires by building type
ggplot(Building_Type_Analysis, aes(x = Type, y = Fires_Per_Building_Perc)) +
  geom_bar(stat = "identity", fill=rgb(0.4,0,0.48627,0.7)) +
  theme_minimal() +
  labs(x = "Building Type",
    y = "Percentage of houses resulting in fires (%)",
```

```

title = "Rate of Fires per Building \n by Building Type") +
theme(plot.title = element_text(hjust = 0.5)) +
coord_cartesian(ylim = c(0,1.5))

```



We can see that in our sample terraced houses tend to have a significantly larger amount of fires associated (**Top Left**). However, by analysing the total amount of each property we find that Utopia has a significantly larger amount of terraced houses than other building types (**Table 1**). We should therefore consider dividing through by the total amount in each group to establish a normalised comparison (**Bottom left**).

When looking at the percentage of houses resulting in fires, we find that terraced houses are still the most likely to be involved in fires, however less significantly than if we just looked at total fires.

It may however be a fair assertion that fires are more likely to occur in terraced houses since there's a higher likelihood of fire spreading house-to-house. While we have a clear outlier comparing the total amount of fires, when adjusting for the amount of each building type, or instead looking at the likelihood of one or more casualties being involved in a fire the relationship between the building type and the other variables is dulled considerably.

We find that no house type is more likely to involve casualties in an incident (**Top Right**), and terraced houses are slightly more likely to be involved in fires than other building types.

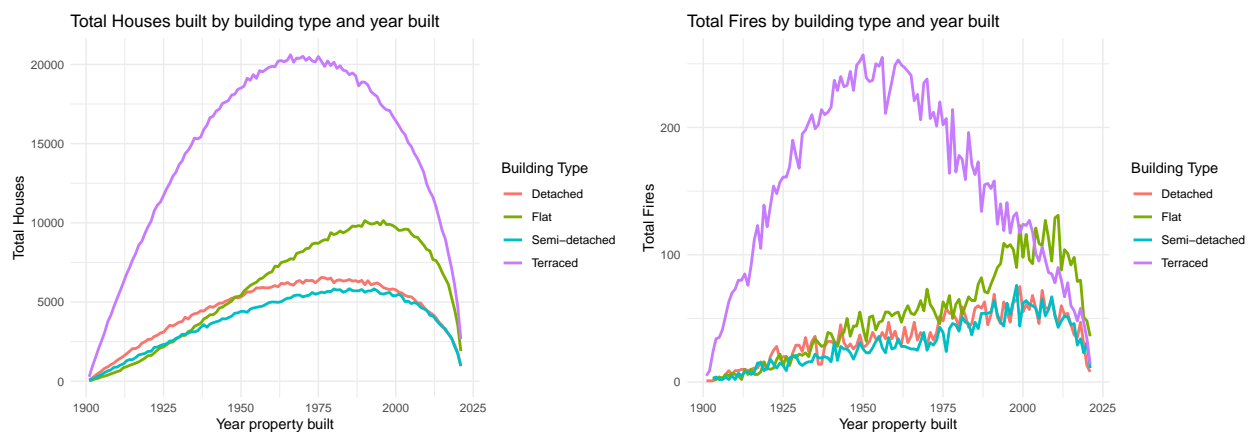
2B - Fires by Building Type

```
# summarise data to find the yearly
Fire_BuildingAge_Analysis = Fire_join %>% group_by(Type,Year) %>%
  summarise(Total_Fires = n())

Years_Houses_Built = Housing_Reg %>% group_by(Type,Year) %>%
  summarise(Total_Houses = n())

# Analyse Houses built by type and year
ggplot(Years_Houses_Built, aes(x = Year, y = Total_Houses, color = Type)) +
  geom_line(linewidth = 1) +
  labs(title = "Total Houses built by building type and year built",
       x = "Year property built",
       y = "Total Houses") +
  scale_color_discrete(name = "Building Type")+
  theme(plot.title = element_text(hjust = 0.5))+
  theme_minimal()

# Analyse Fires by building type and year built
ggplot(Fire_BuildingAge_Analysis, aes(x = Year, y = Total_Fires, color = Type)) +
  geom_line(linewidth = 1) +
  labs(title = "Total Fires by building type and year built",
       x = "Year property built",
       y = "Total Fires") +
  scale_color_discrete(name = "Building Type") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme_minimal()
```



When we analyse fires by year built we find that terraced houses built in the early to late 1900s are involved in a greater number of fires. We must also however analyse the total number of houses built over the same period. We find that the relationship between houses built and fires occurring is near identical, implying that while terraced houses are more likely to be involved in fires, this is only since there are more in the sample.

We can see a downwards trend towards the end of the sample period. It is important to note this does not necessarily mean that newer houses have less fires, since they have had less time for fires to occur in the sample period.

- 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?

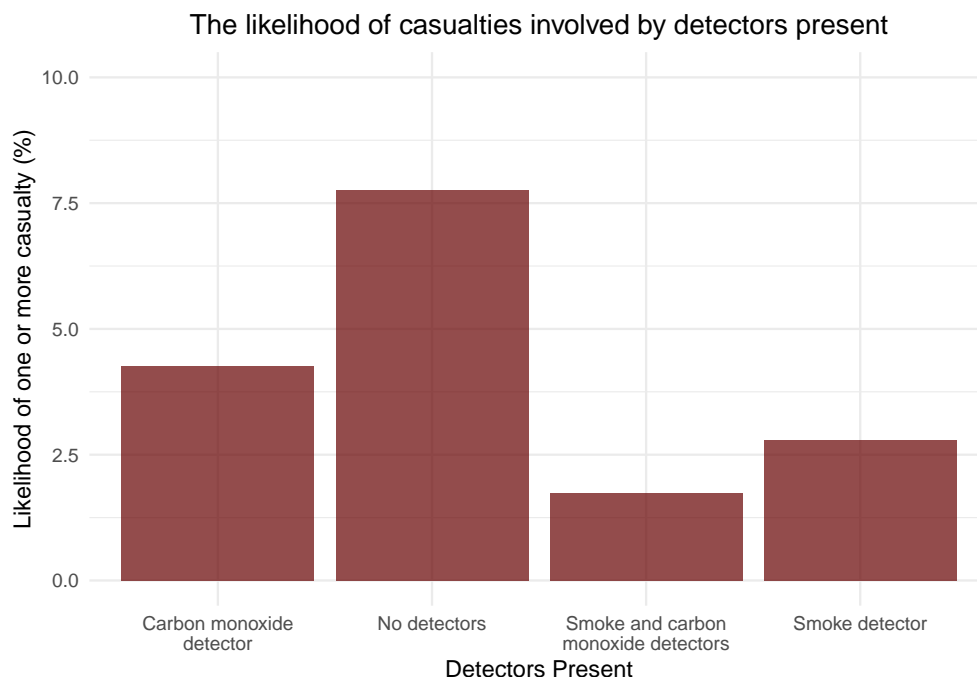
2C - Casualties by Detectors Present

```
# Classify data into groups and summarise

Fires = Fires %>% mutate(Detectors_Present = case_when(
  (Smoke == 1) & (CO == 1) ~ "Smoke and carbon monoxide detectors",
  (Smoke == 1) & (CO == 0) ~ "Smoke detector",
  (Smoke == 0) & (CO == 1) ~ "Carbon monoxide detector",
  (Smoke == 0) & (CO == 0) ~ "No detectors"))

Detector_Analysis = Fires %>%
  group_by(Detectors_Present) %>%
  summarise(Total_Fires = n(),
            Total_Casualties = sum(Casualties),
            Count_Casualties = sum(Casualties > 0),
            Casualty_likelihood = 100*Count_Casualties / Total_Fires,
            Average_Property_Damage = mean(Damage, na.rm = TRUE))

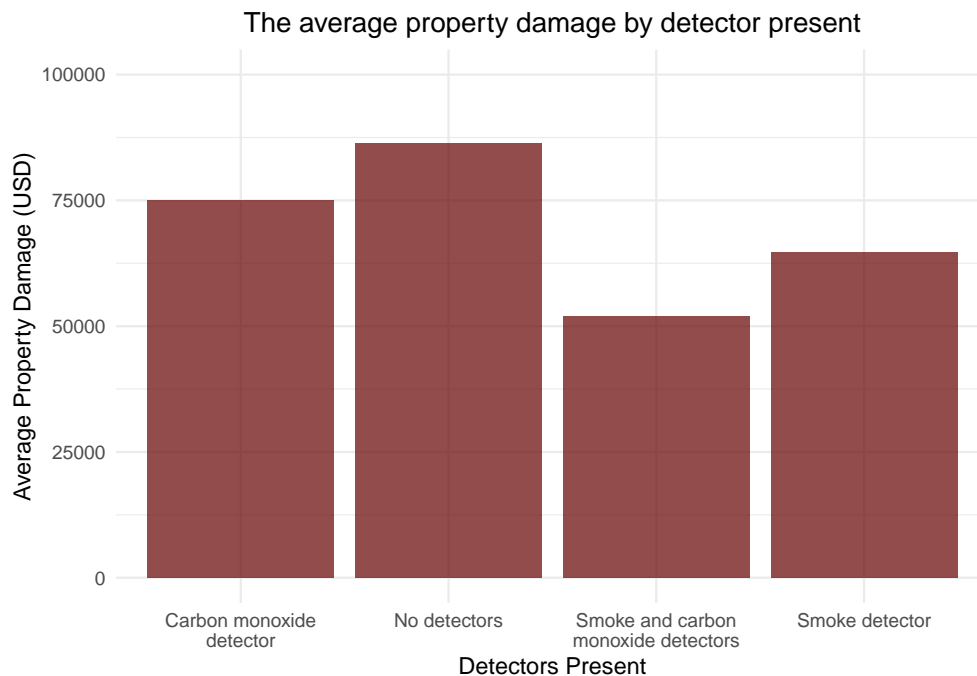
ggplot(Detector_Analysis, aes(x = Detectors_Present, y = Casualty_likelihood)) +
  geom_bar(stat = "identity", fill=rgb(0.4,0,0,0.7)) +
  theme_minimal() +
  labs(x = "Detectors Present",
       y = "Likelihood of one or more casualty (%)",
       title = "The likelihood of casualties involved by detectors present") +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_cartesian(ylim = c(0,10))
```



We analyse that fires where no detectors are present have a significantly higher likelihood in resulting in one or more casualty. The likelihood of casualties is over 4 times higher in properties without detectors.

2C - Property Damage by Detectors Present

```
ggplot(Detector_Analysis, aes(x = Detectors_Present, y = Average_Property_Damage)) +  
  geom_bar(stat = "identity", fill=rgb(0.4,0,0,0.7)) +  
  theme_minimal() +  
  labs(x = "Detectors Present",  
       y = "Average Property Damage (USD)",  
       title = "The average property damage by detector present") +  
  theme(plot.title = element_text(hjust = 0.5)) +  
  coord_cartesian(ylim = c(0,100000))
```



We are able to conclude that properties with no detectors are more likely to have a higher monetary loss than those with detectors. It could be possible that since detectors are able to give early warning of a fire, the emergency services are called sooner and therefore respond quicker than if there were no detector - leading to less damage.