SAD Projekt 1 - Sprawozdanie

Michal Szpunar, Daniel Adamkowski

30 11 2021

Zadanie 1

Przygotowanie danych

Analizę dokonano dla zbioru danych z lipca 2021 roku. Poniżej przedstawiono procedurę wczytania oraz oznaczenia danych:

```
schema <- c(
  'station_id',
  'station_name',
  'year',
  'month',
  'day',
  'max_daily_temp',
  'meas_status_TMAX',
  'min_daily_temp',
  'meas_status_TMIN',
  'mean_daily_temp',
  'meas_status_STD',
  'min_temp_near_ground',
  'meas_status_TMNG',
  'daily_rainfall',
  'meas_status_SMDB',
  'rainfall_type',
  'snow_cover_height',
  'meas_status_PKSN'

)

df <- read.csv('./data/2021_07_k/k_d_07_2021.csv', header=FALSE)
colnames(df) <- schema
```

Następnie stworzono kolumny zawierające różnice pomiędzy maksymalnymi oraz średnimi temperaturami z następujących po sobie dni. W tym celu użyto operacji przekształceń danych dostępnych w bibliotece *dplyr*.

```
library(dplyr)
df <- df %>% group_by(station_name) %>% mutate(max_temp_diff = max_daily_temp - lag(max_daily_temp))
df <- df %>% group_by(station_name) %>% mutate(mean_temp_diff = mean_daily_temp - lag(mean_daily_temp))
```

W celu usunięcia wartości *NaN*, zastąpiono je zerami.

```
df$max_temp_diff[is.na(df$max_temp_diff)] <- 0
df$mean_temp_diff[is.na(df$mean_temp_diff)] <- 0
```

Analiza wybranych punktów pomiarowych

Celem zadania było dokonanie porównania wybranych charakterystyk pogodowych z trzech odległych punktów pomiarowych. Wybrano następujące punkty pomiarowe:

- **PSZCZYNA**
- **GDAŃSK RĘBIECHOWO**
- **WARSZAWA-FILTRY**

Na początek zdecydowano się obliczyć średnie:

- maksymalne dobowe temperatury
- dobowe wahania temperatury
- wahania maksymalnej dobowej temperatury z dnia na dzień

```
mean_temps <- c()
mean_stabilities <- c()
mean_diffs <- c()

for (l in list(loc1, loc2, loc3)){
  mean_temps <- c(mean_temps, mean(l$max_daily_temp))
  mean_stabilities <- c(
    mean_stabilities,
    mean(l$max_daily_temp - l$min_daily_temp)
  )
  mean_diffs <- c(mean_diffs, mean(l$max_temp_diff))
}
```

Otrzymano następujące wyniki (w kolejności PSZCZYNA - GDAŃSK - WARSZAWA)

```
# srednia temperatura w kazdym z punktow pomiarowych
mean_temps
```

```
## [1] 27.01290 25.20323 28.56452
```

```
# srednie dobowe wahania temperatury
mean_stabilities
```

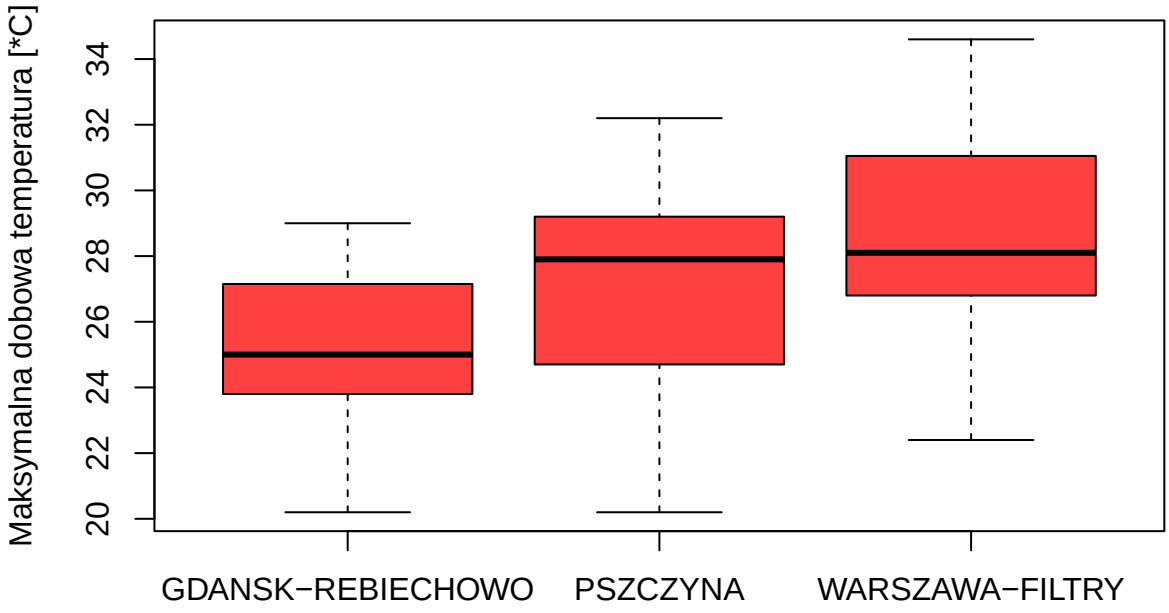
```
## [1] 12.296774  9.496774 10.709677
```

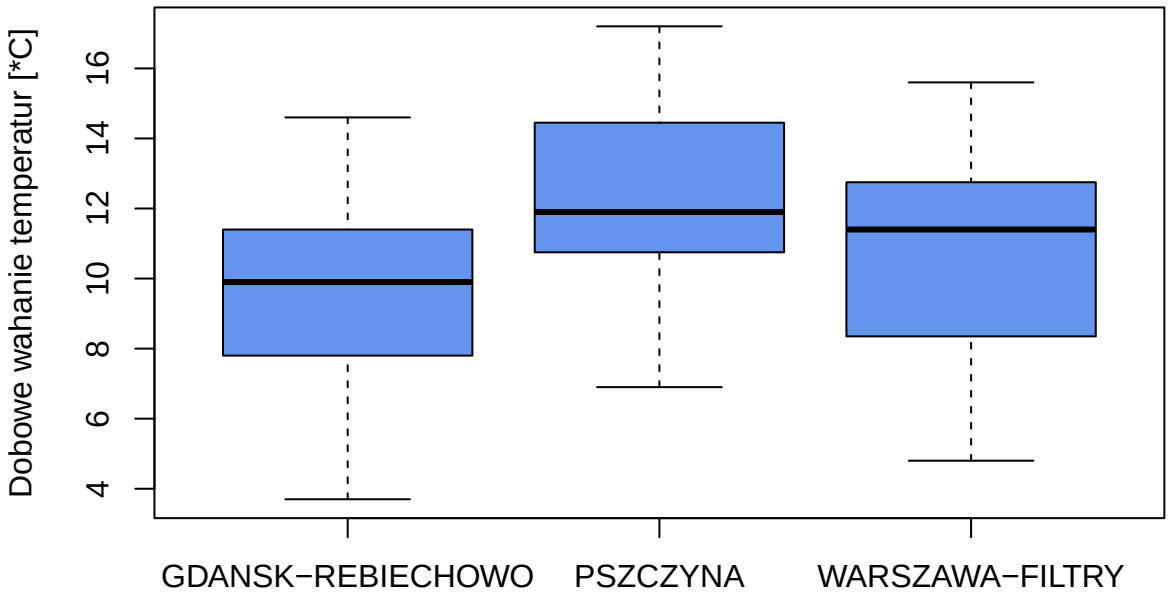
```
# srednie roznice pomiedzy maksymalnymi temp. z dnia na dzien
mean_diffs
```

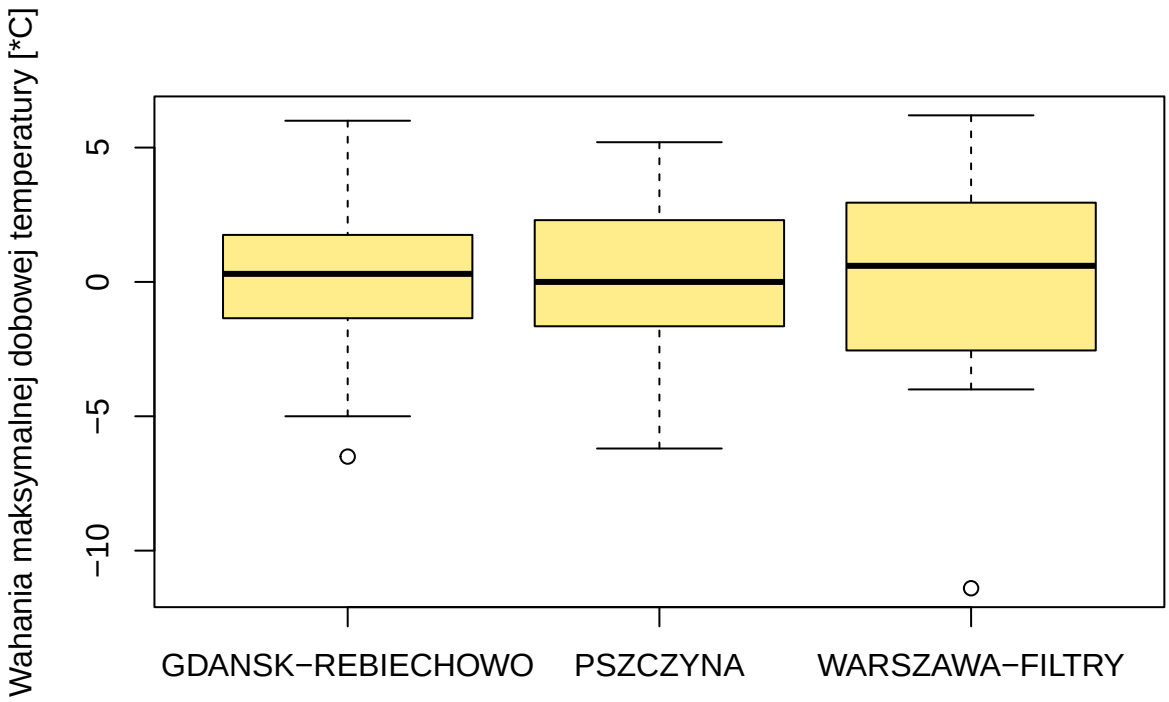
```
## [1] 0.13548387 0.03870968 0.02903226
```

Obliczone statystyki wskazują na to, że średnio w **Warszawie** było **najcieplej**, natomiast **najstabilniejsza dobowa temperatura** była w **Gdańsku**. Aby jednak potwierdzić te hipotezy, stworzono wykresy pudełkowe opisujące w sposób obrazowy zapisane obserwacje.

Wykresy te przedstawiono poniżej:


Maksymalna dobowa temperatura [*C]
20
22
24
26
28
30
32
34
GDANSK−REBIECHOWO
PSZCZYNA
WARSZAWA−FILTRY


Dobowe wahanie temperatur [*C]
4
6
8
10
12
14
16
GDANSK−REBIECHOWO
PSZCZYNA
WARSZAWA−FILTRY



Na podstawie powyższych wykresów można stwierdzić, że:

- rzeczywiście w Warszawie maksymalna temperatura przeważnie jest najwyższa. Najniższa temperatura notowana jest w Gdańsku.
- największa dobowa różnica temperatur występuje w Pszczynie, natomiast najniższa w Gdańsku
- różnice pomiędzy maksymalnymi temperaturami w następujących po sobie dniach we wszystkich stacjach pomiarowych są do siebie zbliżone, zatem ciężko wskazać stację, w której notowania są najwyższe

Badanie rozkładu wahań temperatury "z dnia na dzień"

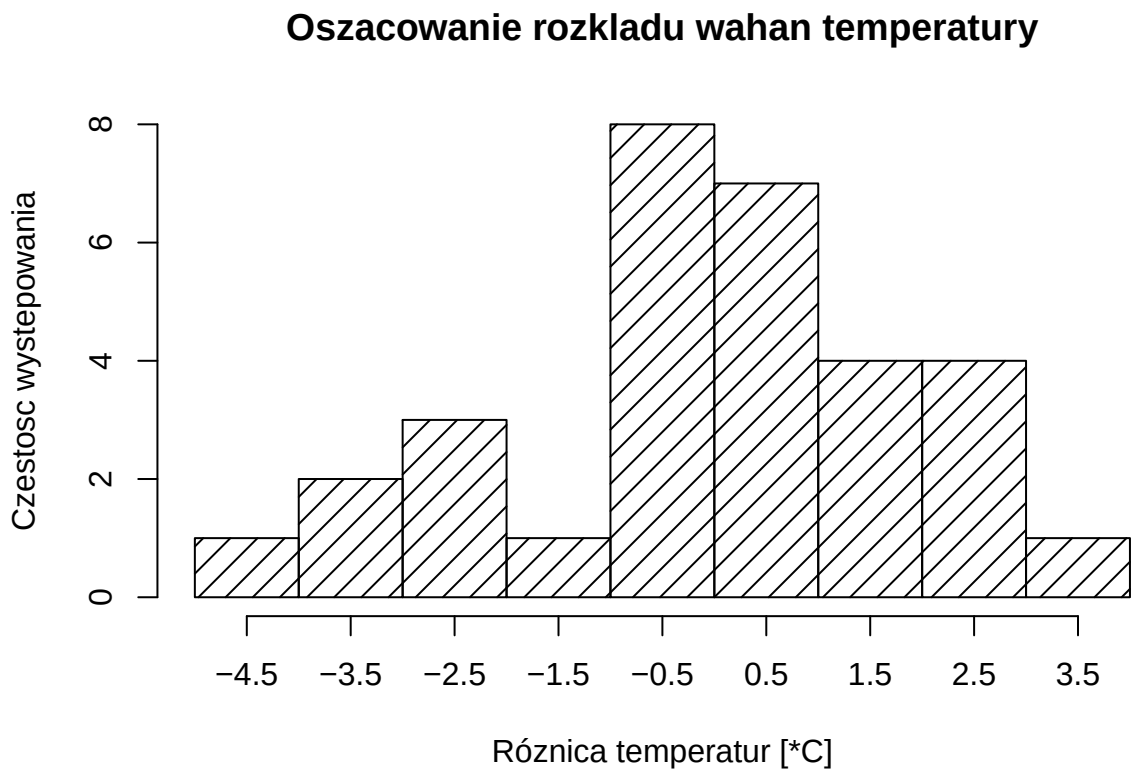
W tej części zadania wybrano stację pomiarową **WARSZAWA-FILTRY**. Na początek stworzono histogram różnic temperatur z dnia na dzień:

```
## Warning in plot.window(xlim, ylim, "", ...): 'grid' nie jest parametrem
## graficznym
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): 'grid'
## nie jest parametrem graficznym
```

```
## Warning in axis(1, ...): 'grid' nie jest parametrem graficznym
```

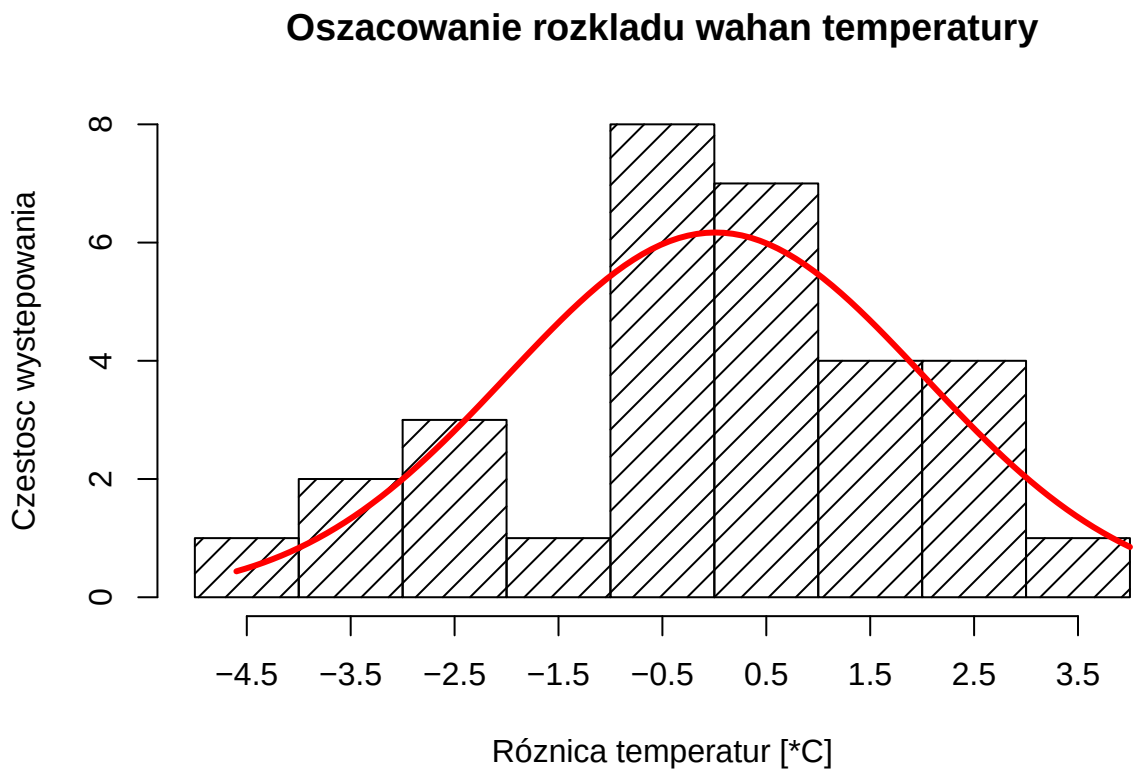
```
## Warning in axis(2, ...): 'grid' nie jest parametrem graficznym
```



Widać, że przypomina kształtem krzywą dzwonową rozkładu Gaussa. Dokonano zatem aproksymacji rozkładu wahań temperatury używając jako modelu rozkładu normalnego. Za wartość oczekiwaną oraz wariancję rozkładu przyjęto średnią oraz wariancję z próby. Na koniec zamieniono funkcję gęstości prawdopodobieństwa na funkcję częstotliwości.

```
linespace <- seq(min(x), max(x), length = 100)
yfit <- dnorm(linespace, mean = mean(x), sd = sd(x))
# converting probability density into frequency function
yfit <- yfit * diff(h$mids[1:2]) * length(x)
```

W wyniku powyższego skryptu otrzymano następującą krzywą:



Po graficznej analizie otrzymanych wyników stwierdzono, że rozkład normalny w dobry sposób oddaje charakterystykę badanej próby.

Według otrzymanego rozkładu prawdopodobieństwa stwierdzenie "jutro będzie tak samo ciepło jak dziś" jest wysoce prawdopodobne.

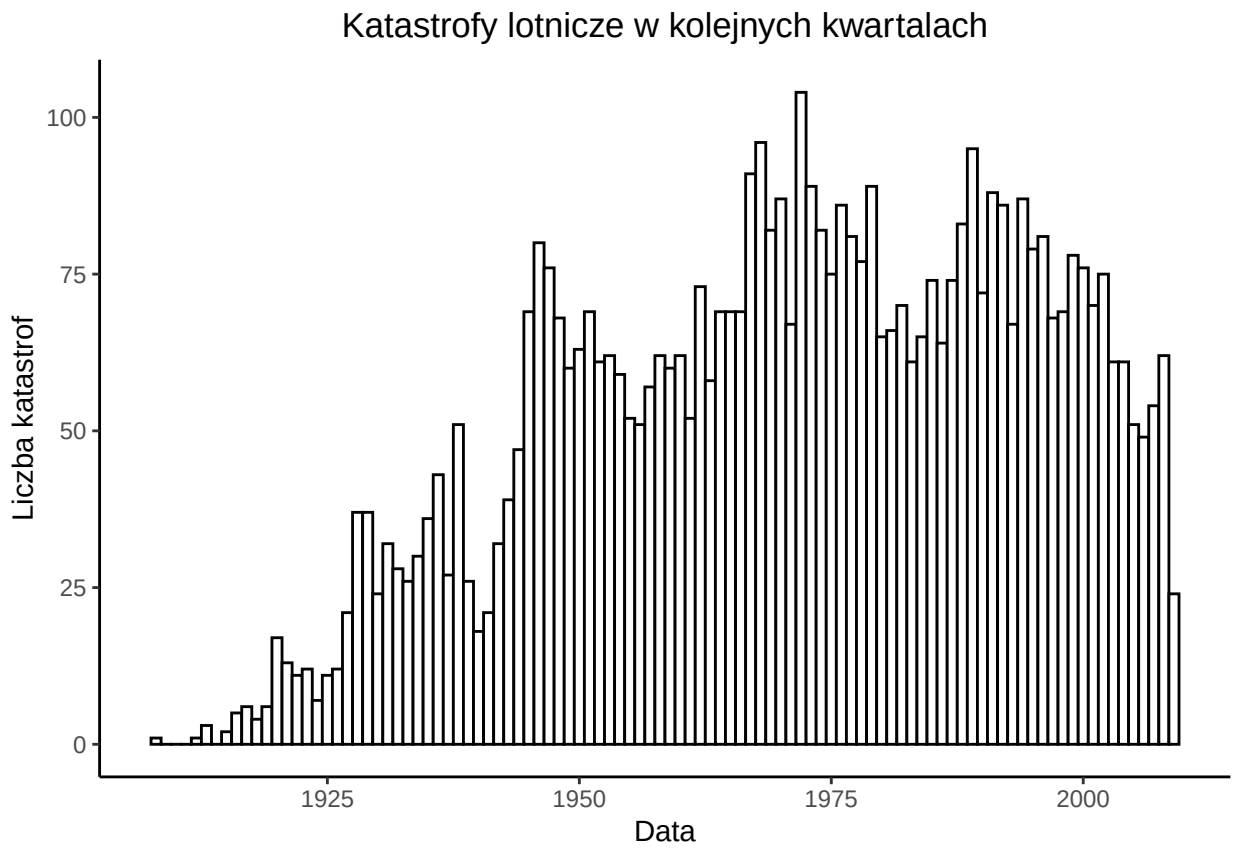
Zadanie 2

a) Badanie liczby katastrof

Przygotowanie danych dotyczących wypadków w kolejnych kwartałach wymagało zakodowania dat w odpowiednim formacie i wyznaczenia na ich podstawie kwartałów.

```
catastrophes <- read.csv(file = "./data/katastrofy.csv")
catastrophes$Date <- as.Date(catastrophes[["Date"]], format = "%m/%d/%Y")
catastrophes$Quarters <- lubridate::quarter(catastrophes$Date, with_year = TRUE)
```

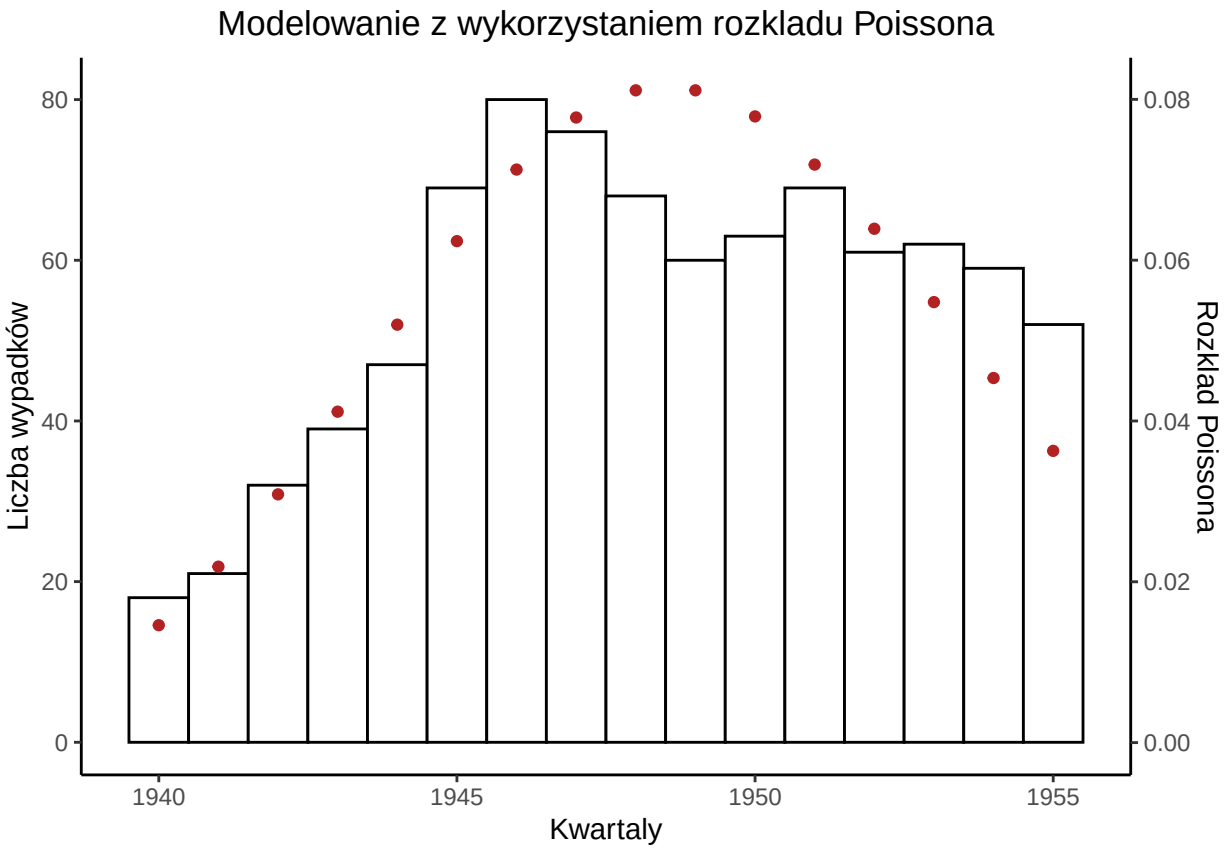
Z przygotowanych danych możemy wygenerować histogram przedstawiający jak zmieniała się czestotliwość zdarzeń w następujących po sobie kwartałach.



Jak widzimy na powyższym wykresie, liczba katastrof lotniczych zmieniała się na przestrzeni lat. Ponadto w związku z długim okresem poddanym analizie oraz ze zmianami zachodzącymi w rzeczywistych procesach związanych z rozwojem lotnictwa, nie jesteśmy w stanie modelować całości danych jednym rozkładem. Jednakowoż możemy wydzielić grupy, pochodzące z krótszych okresów.

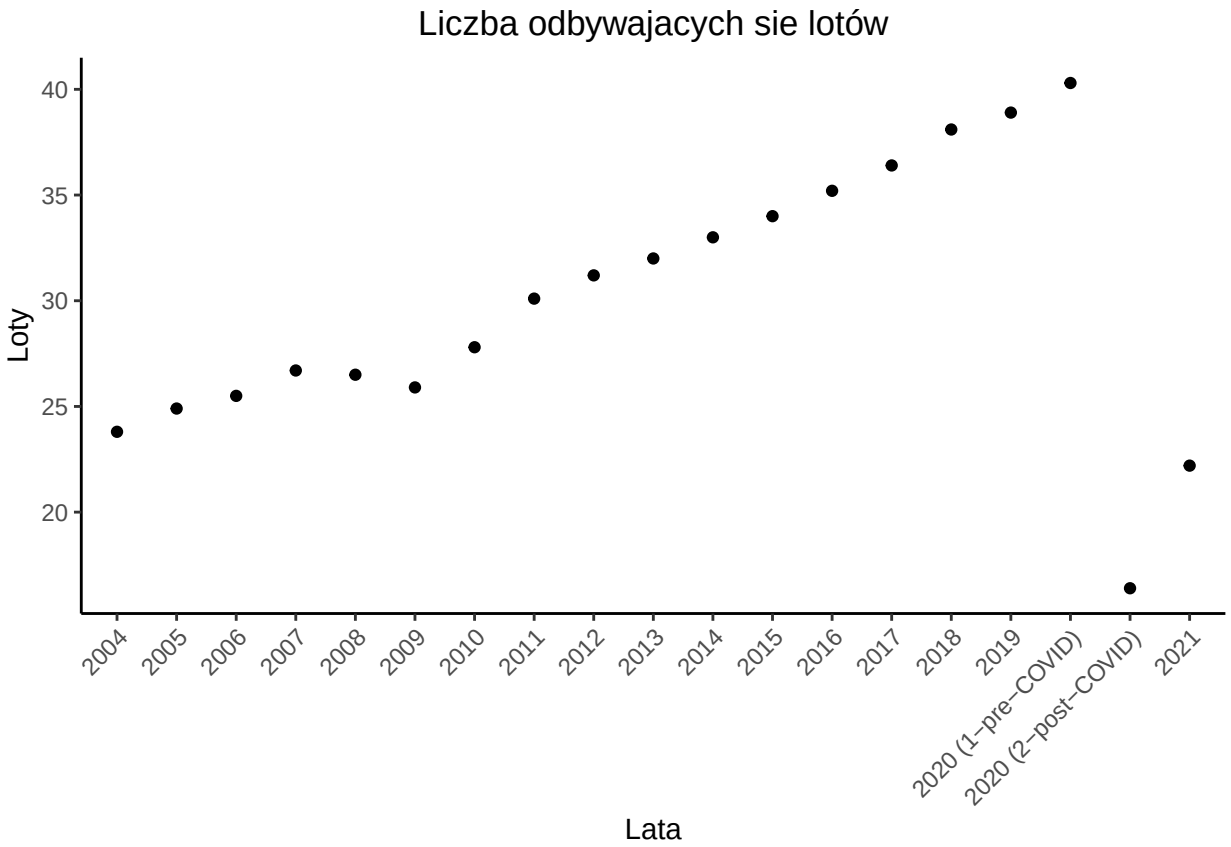
```
cat_subset <- catastrophes[between(catastrophes$Date, as.Date("1940-01-01"), as.Date("1956-01-01")),]

poisson <- as.data.frame(dpois(15:30, lambda = 24))
poisson$x <- 1940:1955
colnames(poisson) <- c("y", "x")
```

Na poniższym wykresie widzimy, że wypadki pochodzące z okresu 1940-1955. Po dobraniu odpowiedniego parametru lambda=24 dla rozkładu Poissona udało się w przybliżeniu zamodelować wybrany fragment danych.



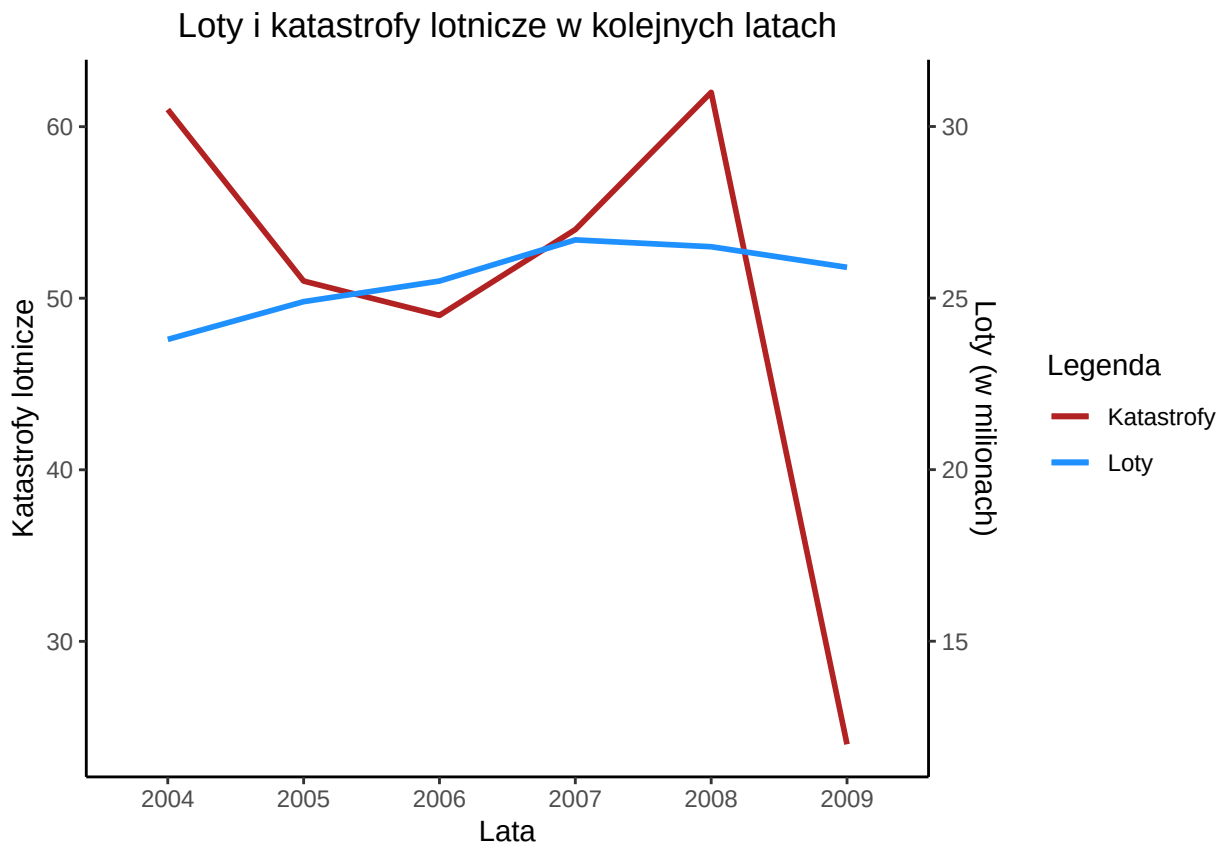
b) Analiza zmiany liczby wypadków na przestrzeni lat

Na podstawie danych pochodzących ze strony https://financesonline.com/number-of-flights-worldwide/ możemy zobaczyć jak zmieniała się ogólna liczba lotów:



Korzystając z faktu, że dla części danych z lat 2004-2009 posiadamy informacje o katastrofach lotniczych, możemy sprawdzić jak zmieniło się bezpieczeństwo lotów.

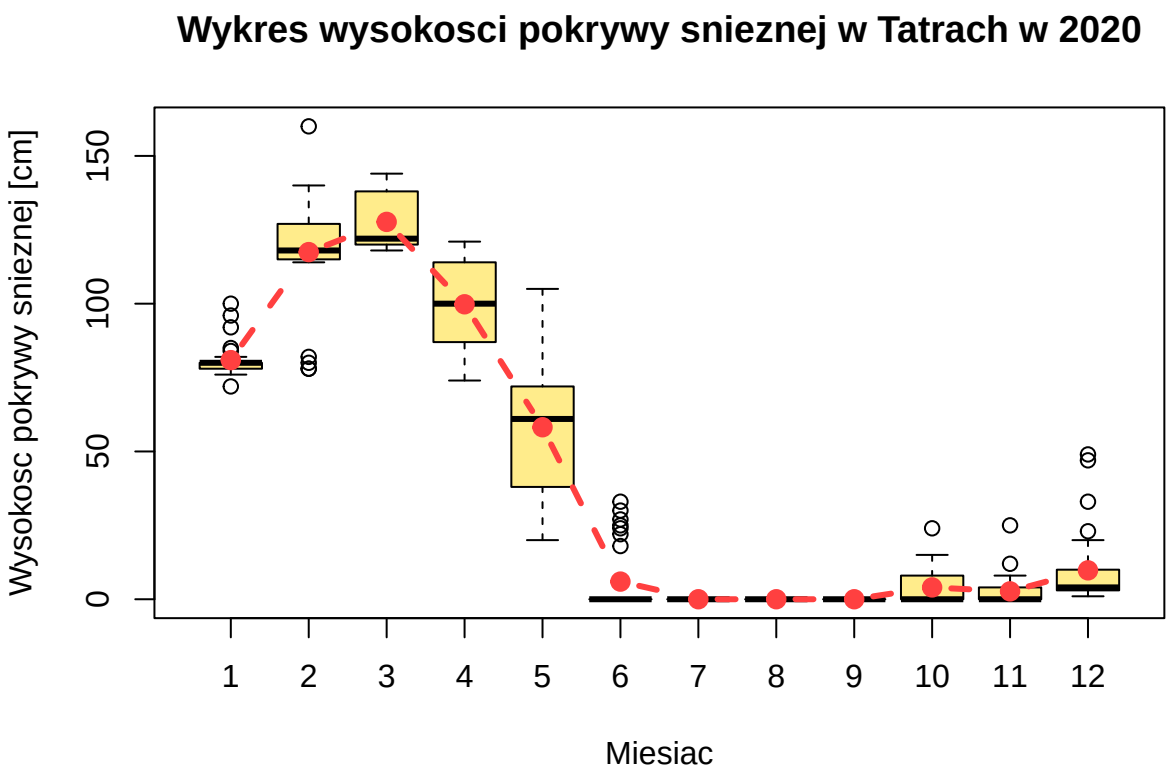
```
years <- lubridate::year(catastrophes$Date)
year_catastrophes <- as.data.frame(table(years[years > 2003]))
year_catastrophes$Flights <- flights[1:6, 2]
colnames(year_catastrophes) <- c("Year", "Catastrophes", "Flights")

ggplot(year_catastrophes, aes(group = 1)) +
  geom_line(aes(x = Year, y = Catastrophes, colour="Katastrofy"), size = 1) +
  geom_line(aes(x = Year, y = Flights * 2, colour="Loty"), size = 1) +
  scale_y_continuous(
    name = "Katastrofy lotnicze",
    sec.axis = sec_axis(trans = ~. * 0.5, name = "Loty (w milionach)")
  ) +
  labs(title="Loty i katastrofy lotnicze w kolejnych latach", x="Lata") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_color_manual(name = "Legenda", values = c("Katastrofy" = "firebrick", "Loty" = "dodgerblue"))
```



Jak widzimy na powyższym wykresie, wzrost ilości lotów nie jest skorelowany z ilością wypadków. W kolejnych latach da się zauważyć wyraźny wzrost bezpieczeństwa.

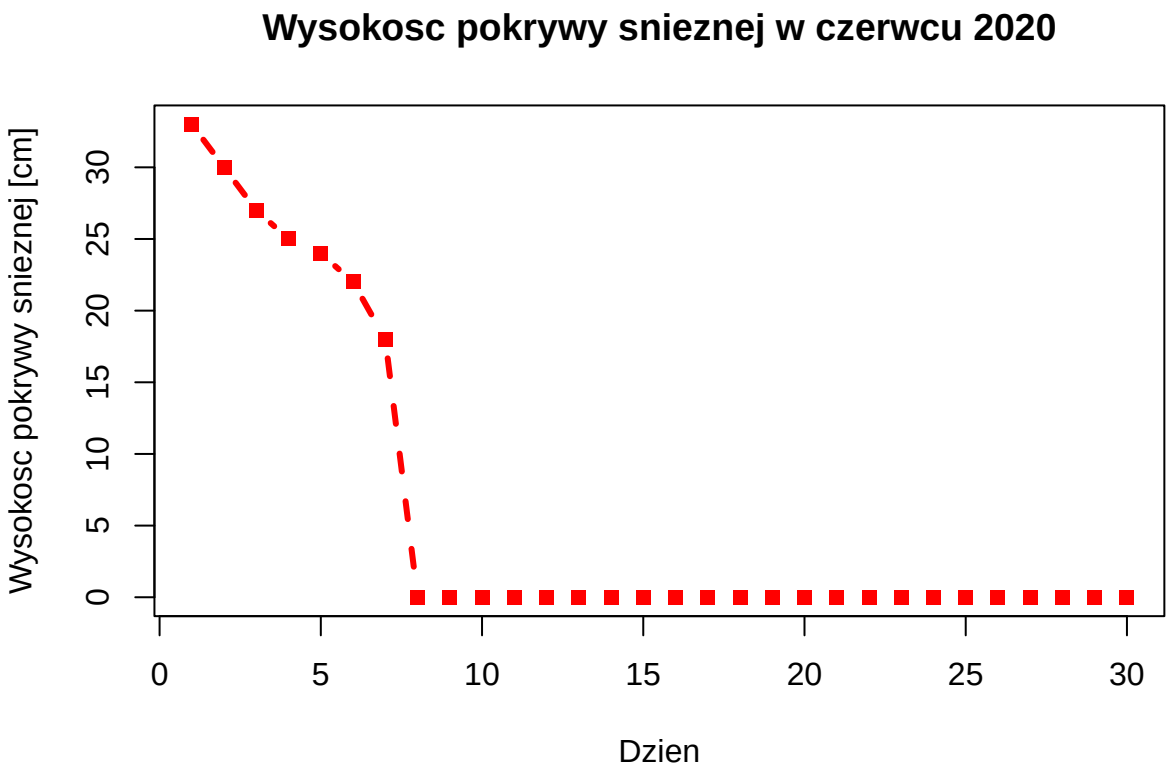
Zadanie 3

Zdecydowano się zbadać wysokość pokrywy śnieżnej w ciągu roku w Tatrzańskim punkcie pomiarowym **Dolina Pięciu Stawów**. W tym celu sporządzono wykres pudełkowy zawierający dane pomiarowe z ostatniego roku. Na wykres naniesiono również wartości średnie dla każdego miesiąca. Wyniki przedstawiono poniżej:



Widać, że w Tatrach śnieg utrzymywał na wysokim poziomie od stycznia do maja. Najwyższe pokrywy odnotowano w marcu, natomiast najniższe od lipca do września - wtedy punkt pomiarowy nie zanotował żadnej pokrywy śnieżnej.

Ze względu na dużą liczbę outlierów, zdecydowano się w szczególny sposób przyjrzeć danym z czerwca. Wyniki przedstawiono na poniższym wykresie:



Powyższy wykres sugeruje, że na początku drugiego tygodnia czerwca śnieg gwałtownie stopniał, zatem okres, w którym nie notowano żadnej pokrywy śnieżnej zaczął się od drugiego tygodnia czerwca i trwał aż do końca września.