SAD - projekt 1

Autorzy:
Konrad Kulesza,
Adam Stec, 300261

Zadanie 1

Zadanie 2

a)

```
if (!require(lubridate))
  install.packages('lubridate')
if (!require(dplyr))
  install.packages('dplyr')

library(lubridate) # for quarter function
library(dplyr)
```

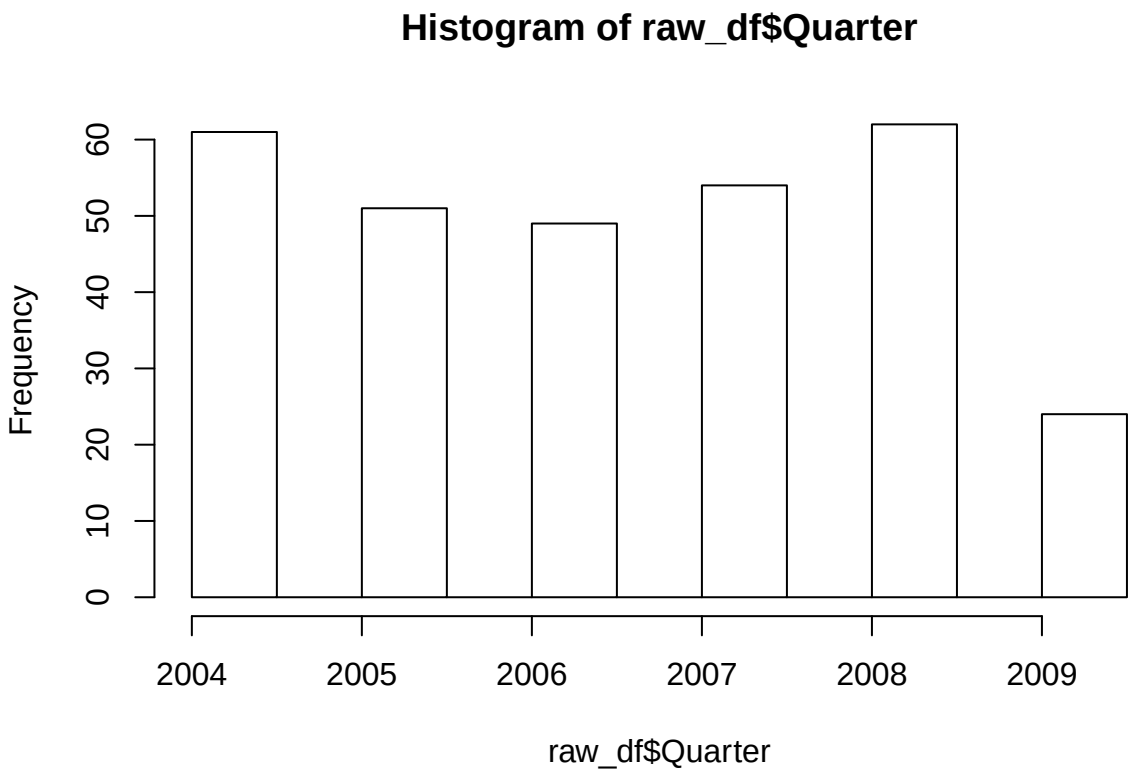
```
raw_df <- read.csv('data/katastrofy.csv', header = TRUE)
raw_df$Date <- as.Date(raw_df$Date,  format= "%m/%d/%Y")
raw_df <- raw_df[raw_df$Date >= "2004-01-01" & raw_df$Date < "2020-01-01", ]


raw_df$Quarter <- quarter(raw_df$Date, with_year = TRUE)
grouped_by_quarter <- raw_df %>% group_by(Quarter) %>% tally()
str(grouped_by_quarter)
```

```
## tibble [22 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Quarter: num [1:22] 2004 2004 2004 2004 2005 ...
##  $ n      : int [1:22] 13 13 15 20 18 10 13 10 12 14 ...
```

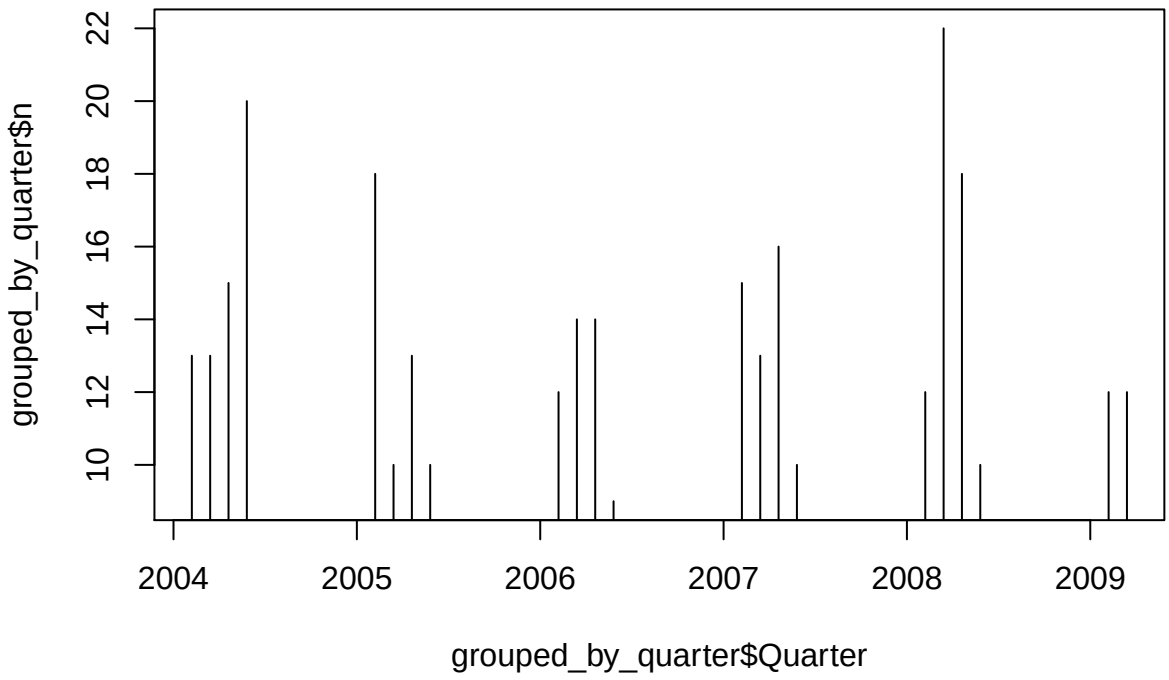
```
mean(grouped_by_quarter$n)
```

```
## [1] 13.68182
```

```
hist(raw_df$Quarter)
```




```
plot(grouped_by_quarter$Quarter, grouped_by_quarter$n, type='h')
```



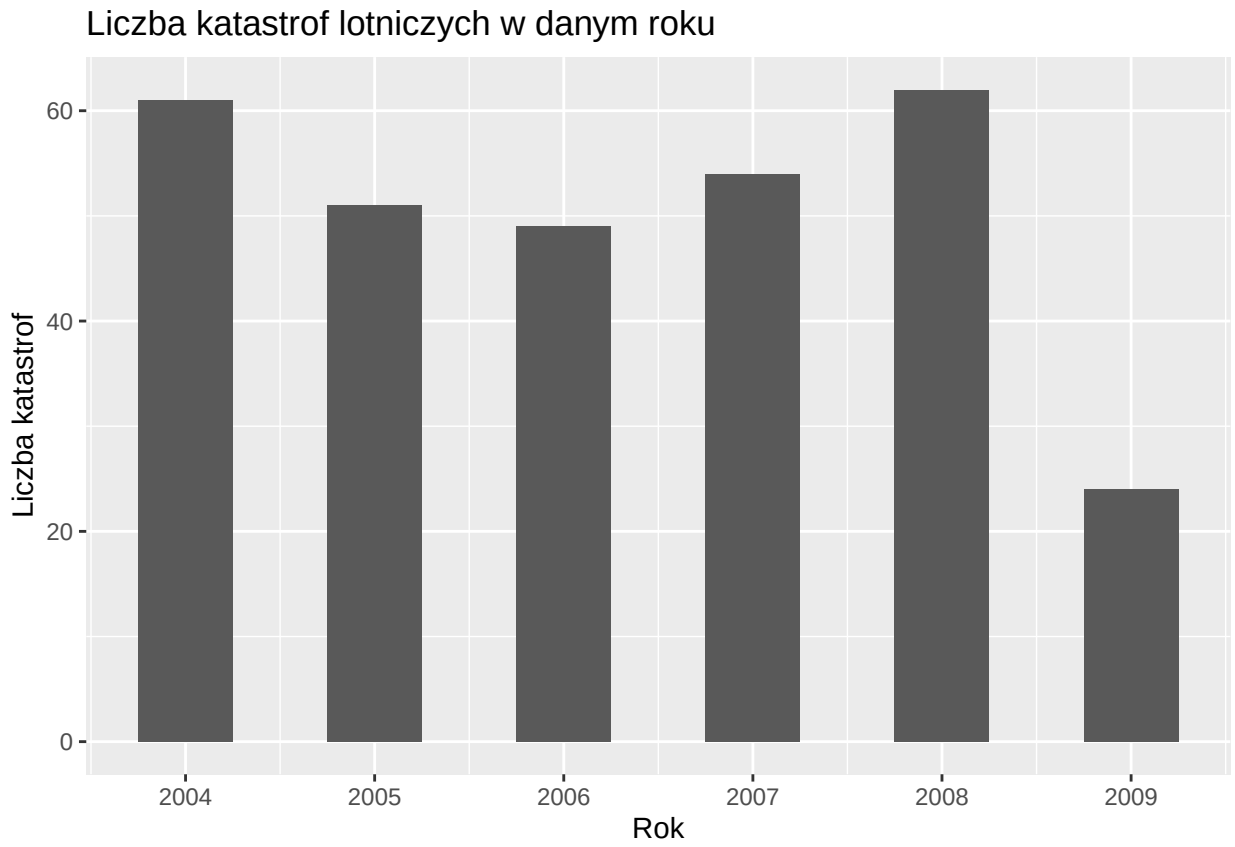
b)

```
require(dplyr)
require(ggplot2)
```

Wykres przedstawiający liczbę katastrof lotniczych w latach 2004 - 2009 (dostępne dane nie zawierają informacji o późniejszych katastrofach)

```
# funkcja pomocnicza
integer_breaks <- function(n = 5, ...) {
  fxn <- function(x) {
    breaks <- floor(pretty(x, n, ...))
    names(breaks) <- attr(breaks, "labels")
    breaks
  }
  return(fxn)
}


tibble(raw_df) -> dt
dt <- dt %>% group_by(year(Date)) %>% tally()
#dt
ggplot(dt, aes(x = `year(Date)`, y = n)) +
  geom_col(width = 0.5) +
  ggtitle("Liczba katastrof lotniczych w danym roku") +
  xlab("Rok") +
  ylab("Liczba katastrof") +
```

```
  scale_x_continuous(breaks = integer_breaks(6))
```



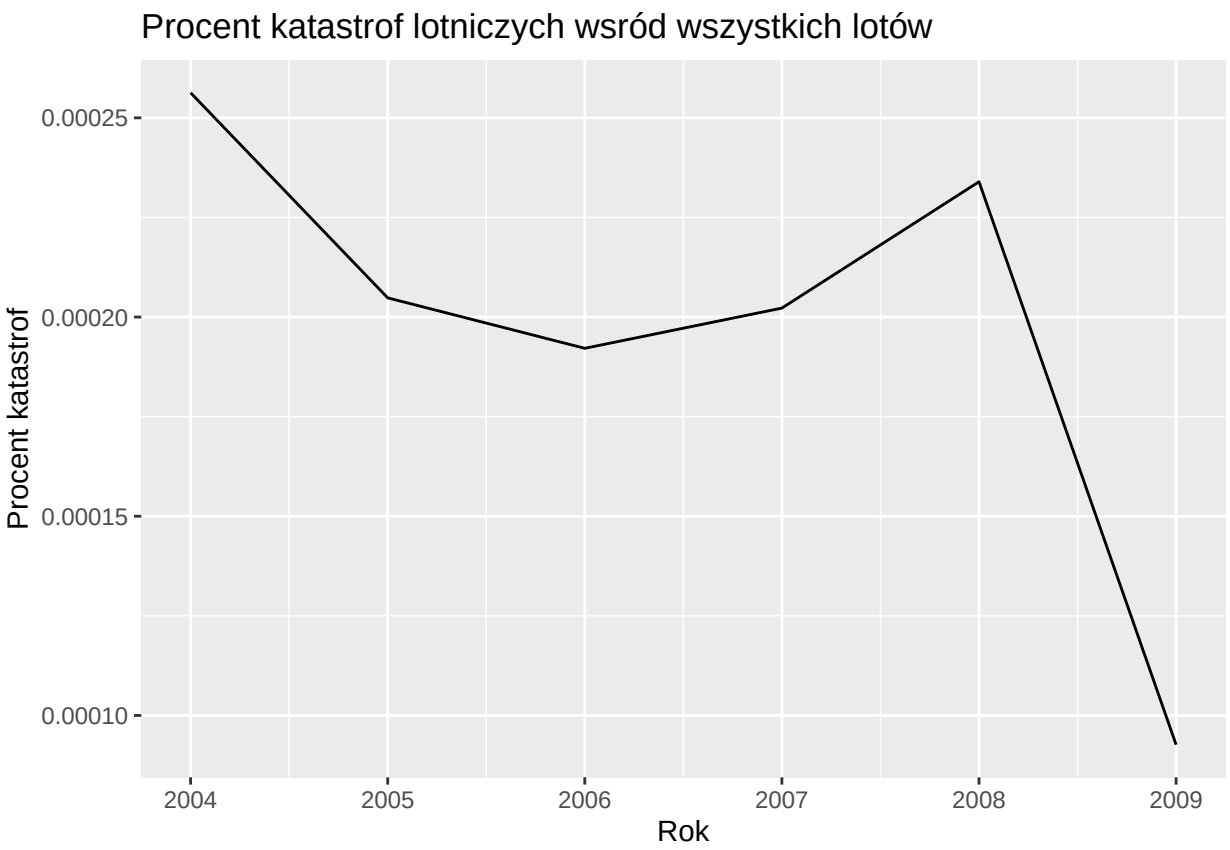
W celu zbadania, ile procent wszystkich odbytych lotów zakończyło się katastrofą, należy połączyć dwie ramki reprezentujące ilość wszystkich lotów oraz katastrof lotniczych dla zadanego roku.

```
num_of_flights <-
  tibble(read.csv('data/number-of-flights-from-2.csv', header = TRUE, nrows = 6))

res <- merge(num_of_flights, dt, by.x = "Category", by.y = "year(Date)")
res$Number.of.Flights.from.2004.to.2021 <-
  res$Number.of.Flights.from.2004.to.2021 * 1000000

res$percentage <- res$n / res$Number.of.Flights.from.2004.to.2021 * 100
res
```

```
##   Category Number.of.Flights.from.2004.to.2021  n   percentage
## 1     2004                             23800000 61 2.563025e-04
## 2     2005                             24900000 51 2.048193e-04
## 3     2006                             25500000 49 1.921569e-04
## 4     2007                             26700000 54 2.022472e-04
## 5     2008                             26500000 62 2.339623e-04
## 6     2009                             25900000 24 9.266409e-05
```

```
ggplot(res, aes(x = Category, y = percentage)) +
  geom_line() +
  ggtitle("Procent katastrof lotniczych wsród wszystkich lotów") +
  xlab("Rok") +
```

```
  ylab("Procent katastrof") +
  scale_x_continuous(breaks = integer_breaks(6))
```



Na podstawie powyższego wykresu można stwierdzić, że na przestrzeni lat 2004 - 2009 loty stały się bezpieczniejsze.

Zadanie 3

Wybrane zostało zbadanie średniej dobowej amplitudy temperatur w poszczególnych miesiącach 2021 roku.

Potrzebne biblioteki:

```
require(ggplot2)
require(tidyverse)
```

Następnie pobieramy wszystkie dane z całego roku do jednej ramki, której dla przejrzystości nadajemy konkretne nazwy.

```
dt <- list.files(pattern = '*csv') %>%
  map_df(~read.csv(., header = F))

names(dt) <- c('station_code', 'station_name', 'year', 'month', 'day',
               'T_MAX', 'status_T_MAX', 'T_MIN', 'status_T_MIN', 'T_AVG', 'status_T_AVG',
               'T_MIN_GROUND','status_T_MIN_GROUND','SUM_OF_PRECIPITATION','status_SUM_OF_PRECIPITATION
               'TYPE_OF_RAINFALL', 'SNOW_COVER_HEIGHT', 'status_SNOW_COVER_HEIGHT')
```

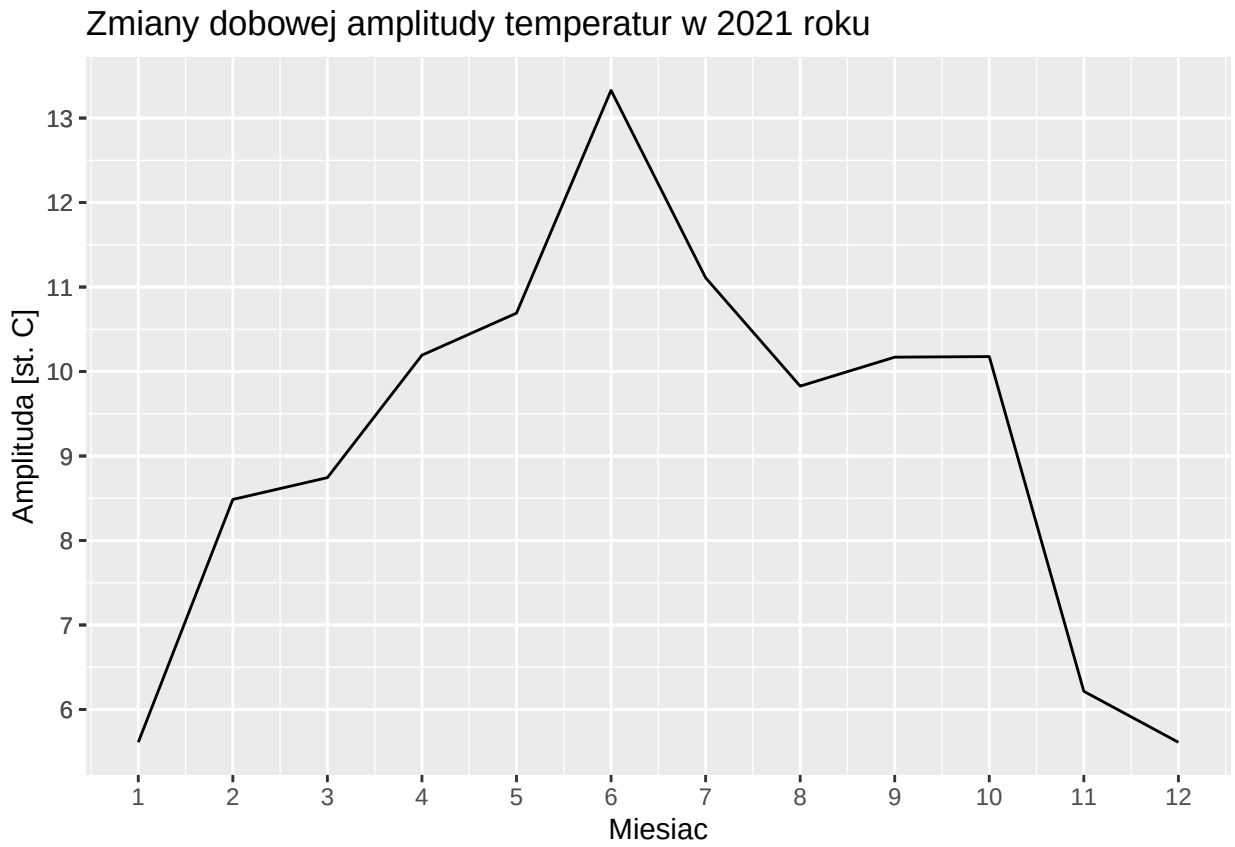
Po tych operacjach należy obliczyć dobową amplitudę, a potem pogrupować wyniki po numerze miesiąca i policzyć dla każdego z nich średnią dobową amplitudę.

```r
dt$amplituda <- dt$T_MAX - dt$T_MIN

dt <- dt %>% group_by(month) %>% summarise(mean_amp = mean(amplituda))

print(dt)
```

```
## # A tibble: 12 x 2
##    month mean_amp
##    <int>    <dbl>
##  1     1     5.61
##  2     2     8.49
##  3     3     8.74
##  4     4    10.2
##  5     5    10.7
##  6     6    13.3
##  7     7    11.1
##  8     8     9.83
##  9     9    10.2
## 10    10    10.2
## 11    11     6.22
## 12    12     5.61
```

Wynik dla lepszego wglądu należy przedstawić na wykresie:

```r
# funkcja pomocnicza
integer_breaks <- function(n = 5, ...) {
  fxn <- function(x) {
    breaks <- floor(pretty(x, n, ...))
    names(breaks) <- attr(breaks, "labels")
    breaks
  }
  return(fxn)
}

ggplot(dt, aes(x = month, y = mean_amp)) +
  geom_line() +
  ggtitle("Zmiany dobowej amplitudy temperatur w 2021 roku") +
  xlab("Miesiac") +
  ylab("Amplituda [st. C]") +
  scale_x_continuous(breaks = integer_breaks(n = 12)) +
  scale_y_continuous(breaks = integer_breaks(n = 13))
```


Zmiany dobowej amplitudy temperatur w 2021 roku
Amplituda [st. C]
13
12
11
10
9
8
7
6
1
2
3
4
5
6
7
8
9
10
11
12
Miesiac