Projekt grupowy PDU

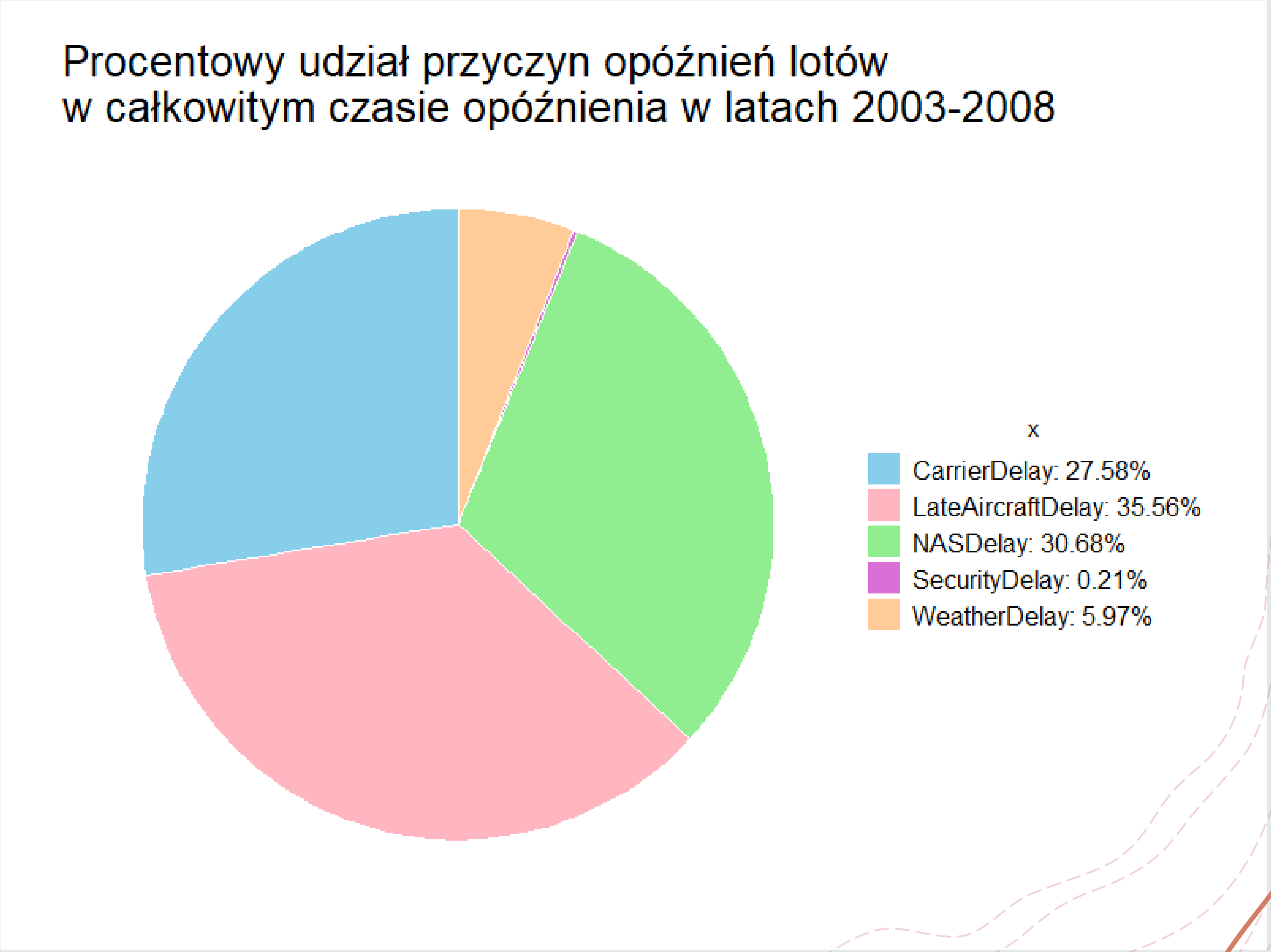
Julia Dollani, Aleksandra Wójcik, Jan Opala

ANALIZA WPŁYWU RÓŻNYCH CZYNNIKÓW NA OPÓŹNIENIA LOTÓW

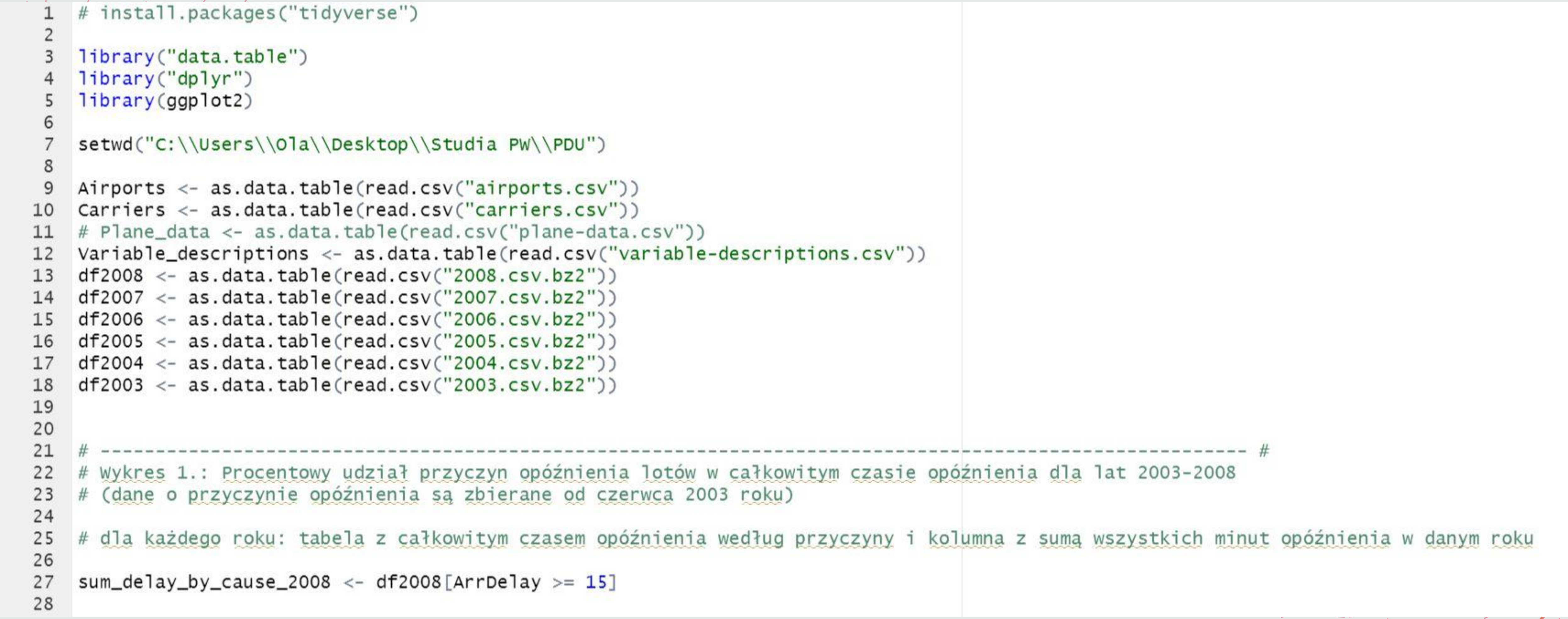
Kategorie:

- Przewoźnik lotniczy
- Ekstremalne warunki pogodowe
- Krajowy system lotniczy USA (NAS – National Aviation System)
- Spóźniony poprzedni samolot
- Bezpieczeństwo

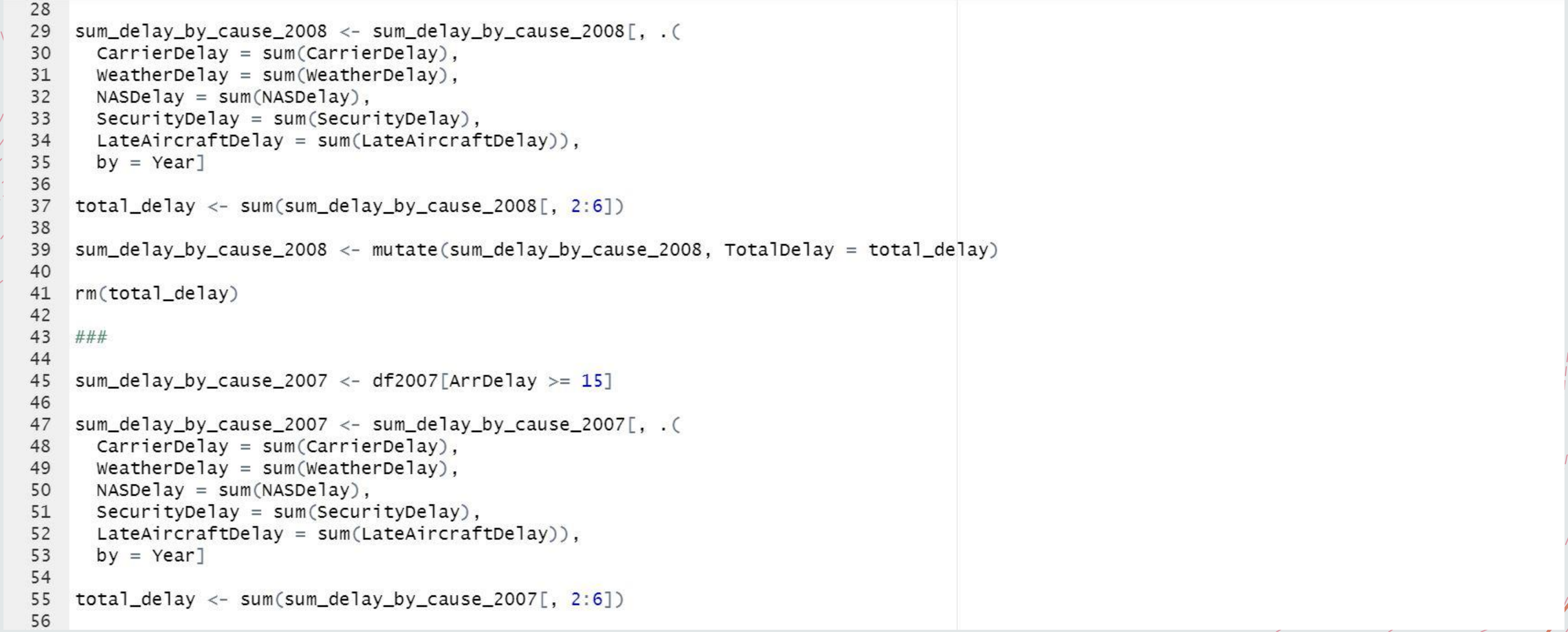
Dane o przyczynach opóźnień zbierane są od czerwca 2003 roku.



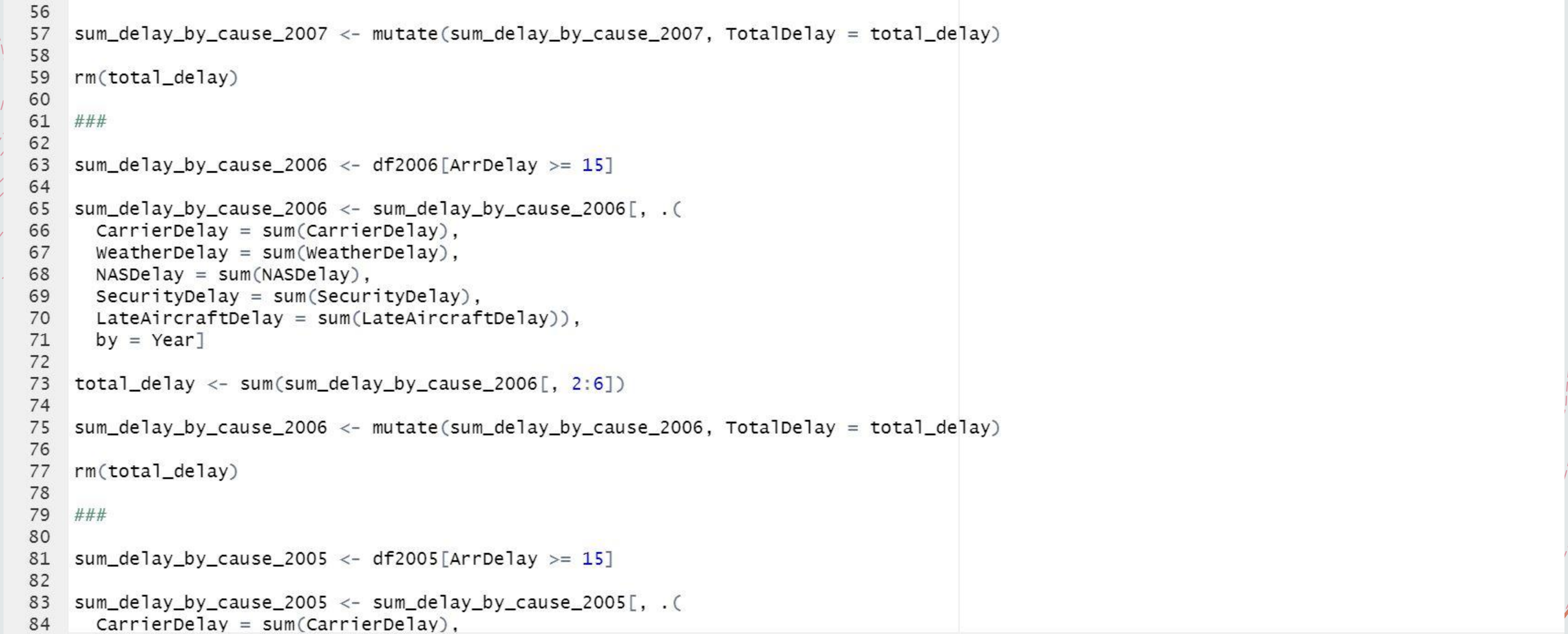
Procentowy udział przyczyn opóźnień lotów w całkowitym czasie opóźnienia w latach 2003-2008



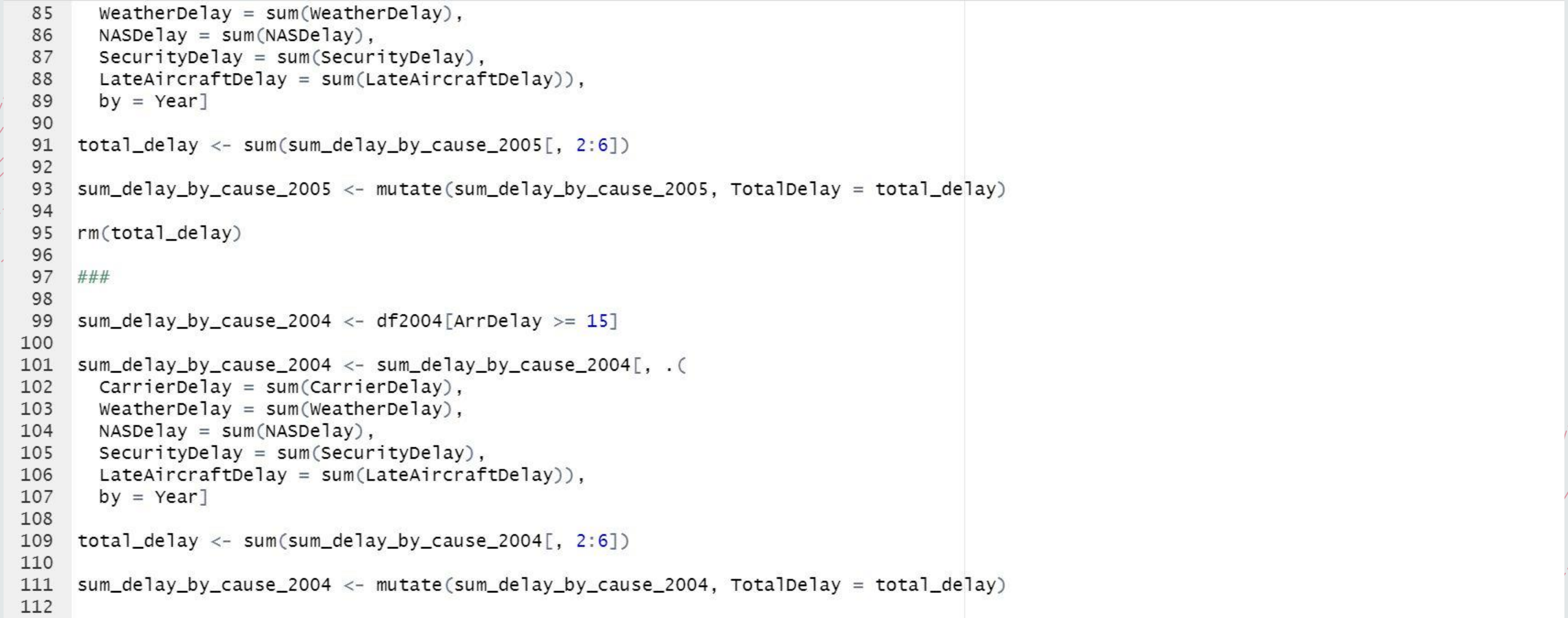
```r
# install.packages("tidyverse")

library("data.table")
library("dplyr")
library(ggplot2)

setwd("C:\\Users\\Ola\\Desktop\\Studia PW\\PDU")

Airports <- as.data.table(read.csv("airports.csv"))
Carriers <- as.data.table(read.csv("carriers.csv"))
# Plane_data <- as.data.table(read.csv("plane-data.csv"))
Variable_descriptions <- as.data.table(read.csv("variable-descriptions.csv"))
df2008 <- as.data.table(read.csv("2008.csv.bz2"))
df2007 <- as.data.table(read.csv("2007.csv.bz2"))
df2006 <- as.data.table(read.csv("2006.csv.bz2"))
df2005 <- as.data.table(read.csv("2005.csv.bz2"))
df2004 <- as.data.table(read.csv("2004.csv.bz2"))
df2003 <- as.data.table(read.csv("2003.csv.bz2"))


# --------------------------------------------------------------------------------------------------------- #
# Wykres 1.: Procentowy udział przyczyn opóźnienia lotów w całkowitym czasie opóźnienia dla lat 2003-2008
# (dane o przyczynie opóźnienia są zbierane od czerwca 2003 roku)

# dla każdego roku: tabela z całkowitym czasem opóźnienia według przyczyny i kolumna z sumą wszystkich minut opóźnienia w danym roku

sum_delay_by_cause_2008 <- df2008[ArrDelay >= 15]

```



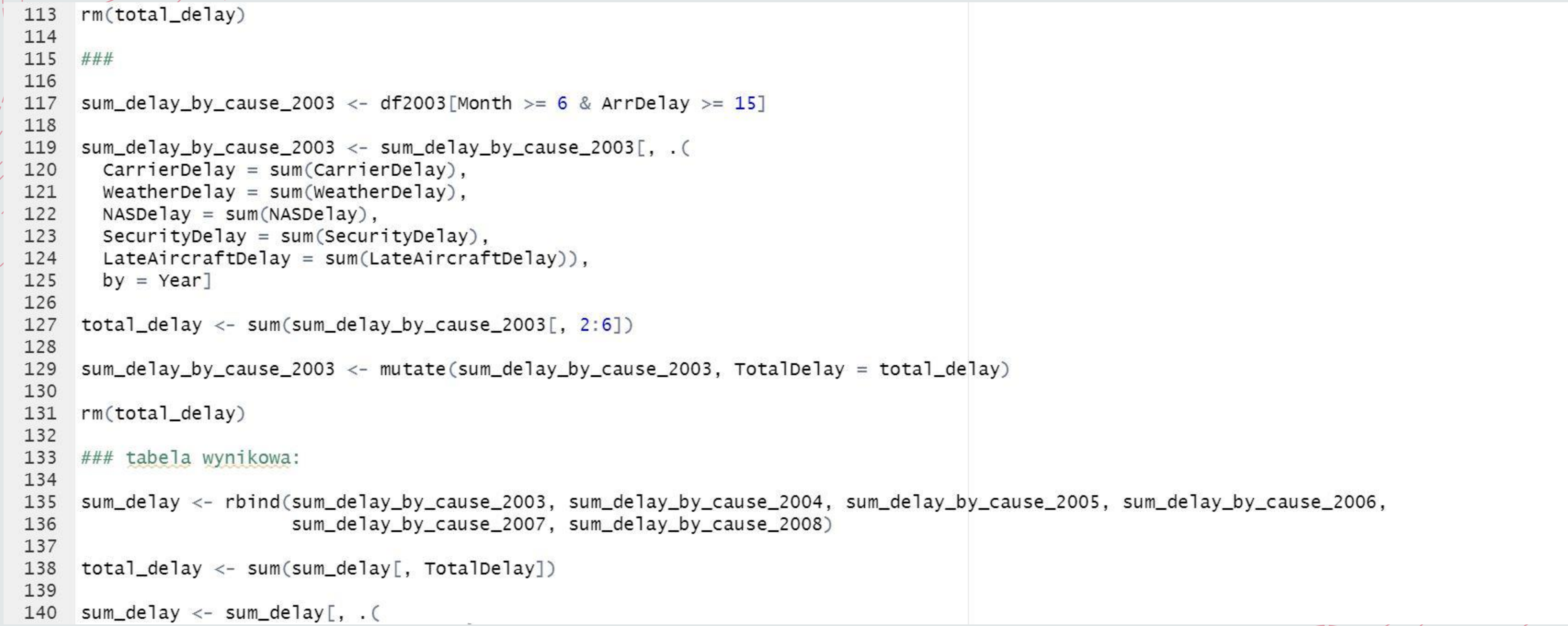
```r

sum_delay_by_cause_2008 <- sum_delay_by_cause_2008[, .(
  CarrierDelay = sum(CarrierDelay),
  WeatherDelay = sum(WeatherDelay),
  NASDelay = sum(NASDelay),
  SecurityDelay = sum(SecurityDelay),
  LateAircraftDelay = sum(LateAircraftDelay)),
  by = Year]

total_delay <- sum(sum_delay_by_cause_2008[, 2:6])

sum_delay_by_cause_2008 <- mutate(sum_delay_by_cause_2008, TotalDelay = total_delay)

rm(total_delay)

###

sum_delay_by_cause_2007 <- df2007[ArrDelay >= 15]

sum_delay_by_cause_2007 <- sum_delay_by_cause_2007[, .(
  CarrierDelay = sum(CarrierDelay),
  WeatherDelay = sum(WeatherDelay),
  NASDelay = sum(NASDelay),
  SecurityDelay = sum(SecurityDelay),
  LateAircraftDelay = sum(LateAircraftDelay)),
  by = Year]

total_delay <- sum(sum_delay_by_cause_2007[, 2:6])

```



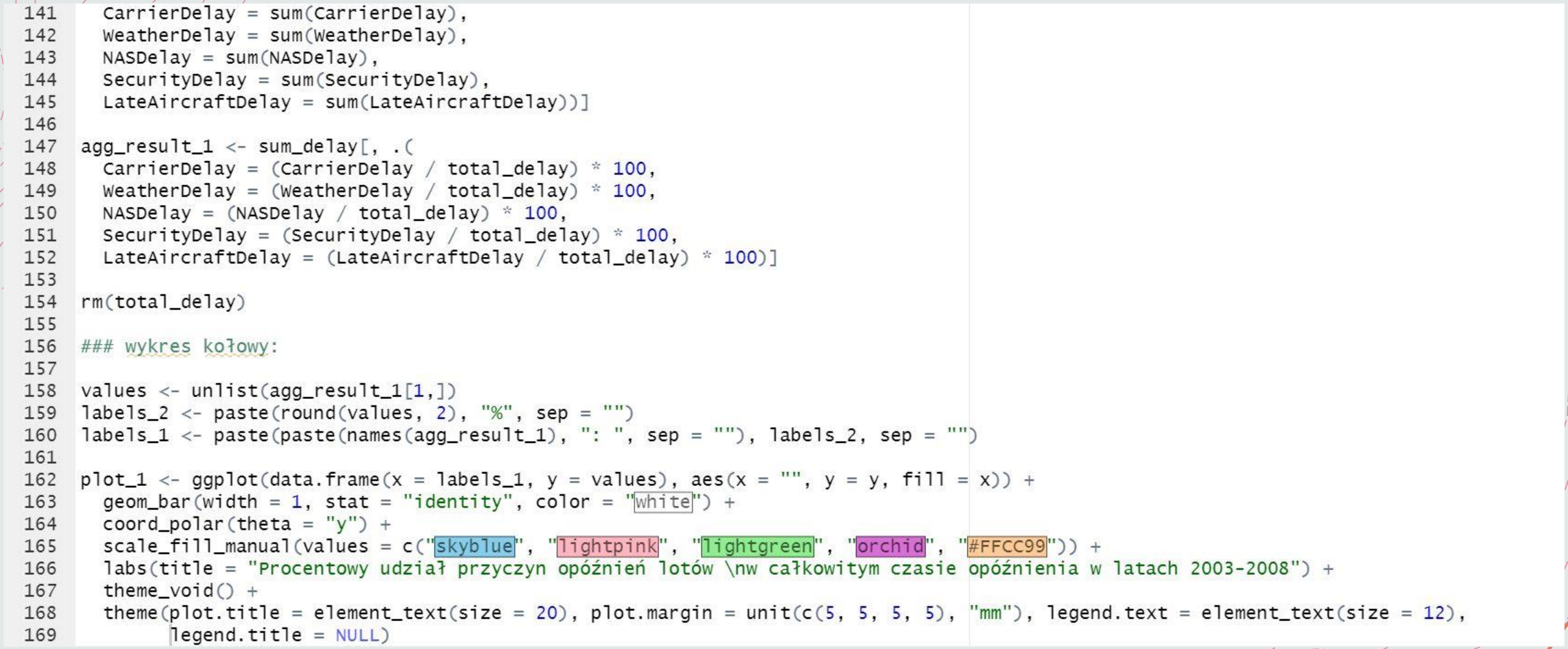
```r

sum_delay_by_cause_2007 <- mutate(sum_delay_by_cause_2007, TotalDelay = total_delay)

rm(total_delay)

###

sum_delay_by_cause_2006 <- df2006[ArrDelay >= 15]

sum_delay_by_cause_2006 <- sum_delay_by_cause_2006[, .(
  CarrierDelay = sum(CarrierDelay),
  WeatherDelay = sum(WeatherDelay),
  NASDelay = sum(NASDelay),
  SecurityDelay = sum(SecurityDelay),
  LateAircraftDelay = sum(LateAircraftDelay)),
  by = Year]

total_delay <- sum(sum_delay_by_cause_2006[, 2:6])

sum_delay_by_cause_2006 <- mutate(sum_delay_by_cause_2006, TotalDelay = total_delay)

rm(total_delay)

###

sum_delay_by_cause_2005 <- df2005[ArrDelay >= 15]

sum_delay_by_cause_2005 <- sum_delay_by_cause_2005[, .(
  CarrierDelay = sum(CarrierDelay),
```



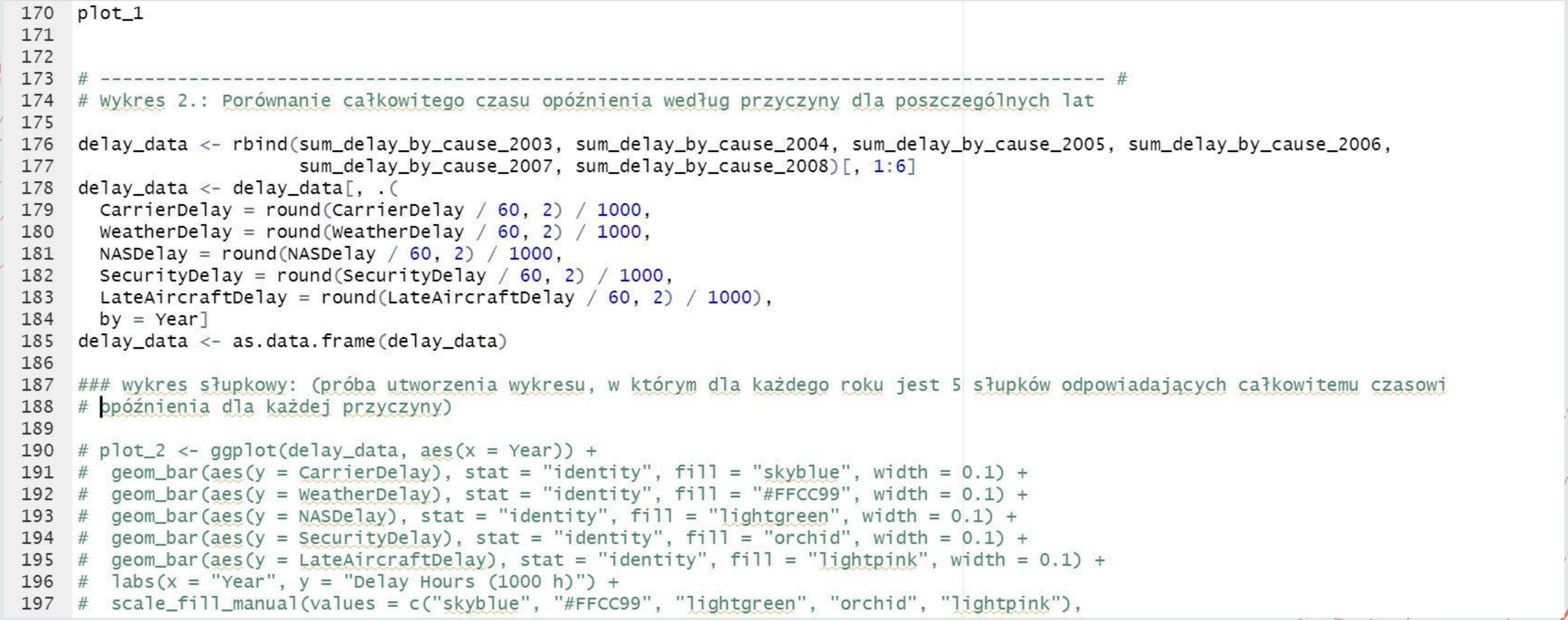
```
  WeatherDelay = sum(WeatherDelay),
  NASDelay = sum(NASDelay),
  SecurityDelay = sum(SecurityDelay),
  LateAircraftDelay = sum(LateAircraftDelay)),
  by = Year]

total_delay <- sum(sum_delay_by_cause_2005[, 2:6])

sum_delay_by_cause_2005 <- mutate(sum_delay_by_cause_2005, TotalDelay = total_delay)

rm(total_delay)

###

sum_delay_by_cause_2004 <- df2004[ArrDelay >= 15]

sum_delay_by_cause_2004 <- sum_delay_by_cause_2004[, .(
  CarrierDelay = sum(CarrierDelay),
  WeatherDelay = sum(WeatherDelay),
  NASDelay = sum(NASDelay),
  SecurityDelay = sum(SecurityDelay),
  LateAircraftDelay = sum(LateAircraftDelay)),
  by = Year]

total_delay <- sum(sum_delay_by_cause_2004[, 2:6])

sum_delay_by_cause_2004 <- mutate(sum_delay_by_cause_2004, TotalDelay = total_delay)

```



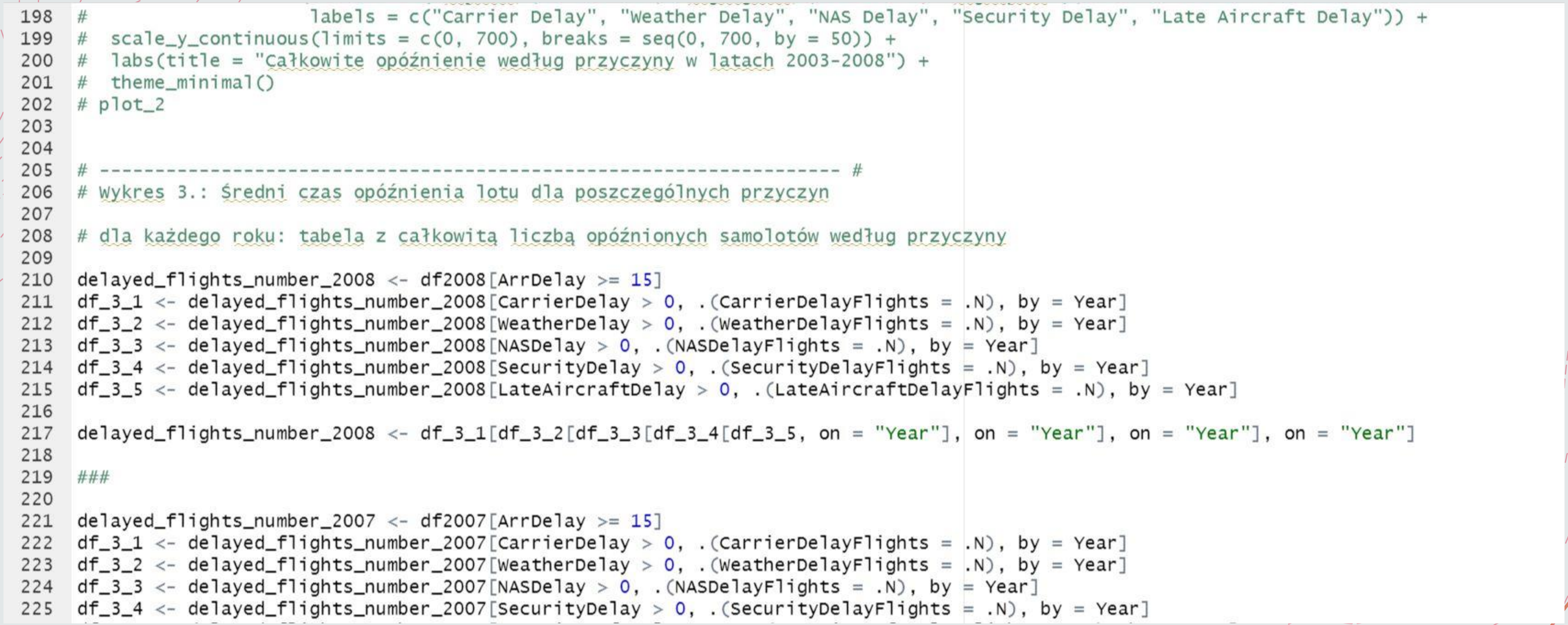
```r
rm(total_delay)

###

sum_delay_by_cause_2003 <- df2003[Month >= 6 & ArrDelay >= 15]

sum_delay_by_cause_2003 <- sum_delay_by_cause_2003[, .(
  CarrierDelay = sum(CarrierDelay),
  WeatherDelay = sum(WeatherDelay),
  NASDelay = sum(NASDelay),
  SecurityDelay = sum(SecurityDelay),
  LateAircraftDelay = sum(LateAircraftDelay)),
  by = Year]

total_delay <- sum(sum_delay_by_cause_2003[, 2:6])

sum_delay_by_cause_2003 <- mutate(sum_delay_by_cause_2003, TotalDelay = total_delay)

rm(total_delay)

### tabela wynikowa:

sum_delay <- rbind(sum_delay_by_cause_2003, sum_delay_by_cause_2004, sum_delay_by_cause_2005, sum_delay_by_cause_2006,
                   sum_delay_by_cause_2007, sum_delay_by_cause_2008)

total_delay <- sum(sum_delay[, TotalDelay])

sum_delay <- sum_delay[, .(
```



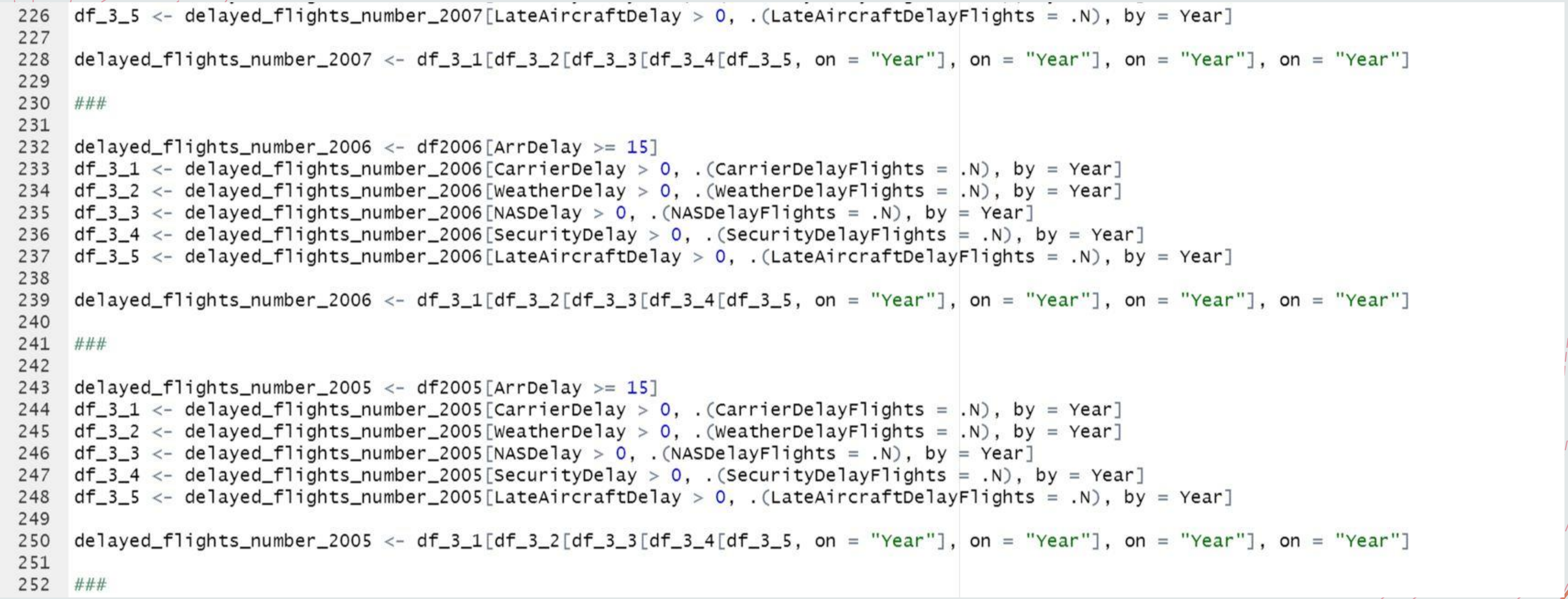
```r
  CarrierDelay = sum(CarrierDelay),
  WeatherDelay = sum(WeatherDelay),
  NASDelay = sum(NASDelay),
  SecurityDelay = sum(SecurityDelay),
  LateAircraftDelay = sum(LateAircraftDelay))]

agg_result_1 <- sum_delay[, .(
  CarrierDelay = (CarrierDelay / total_delay) * 100,
  WeatherDelay = (WeatherDelay / total_delay) * 100,
  NASDelay = (NASDelay / total_delay) * 100,
  SecurityDelay = (SecurityDelay / total_delay) * 100,
  LateAircraftDelay = (LateAircraftDelay / total_delay) * 100)]

rm(total_delay)

### wykres kołowy:

values <- unlist(agg_result_1[1,])
labels_2 <- paste(round(values, 2), "%", sep = "")
labels_1 <- paste(paste(names(agg_result_1), ": ", sep = ""), labels_2, sep = "")

plot_1 <- ggplot(data.frame(x = labels_1, y = values), aes(x = "", y = y, fill = x)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar(theta = "y") +
  scale_fill_manual(values = c("skyblue", "lightpink", "lightgreen", "orchid", "#FFCC99")) +
  labs(title = "Procentowy udział przyczyn opóźnień lotów \nw całkowitym czasie opóźnienia w latach 2003-2008") +
  theme_void() +
  theme(plot.title = element_text(size = 20), plot.margin = unit(c(5, 5, 5, 5), "mm"), legend.text = element_text(size = 12),
        legend.title = NULL)
```



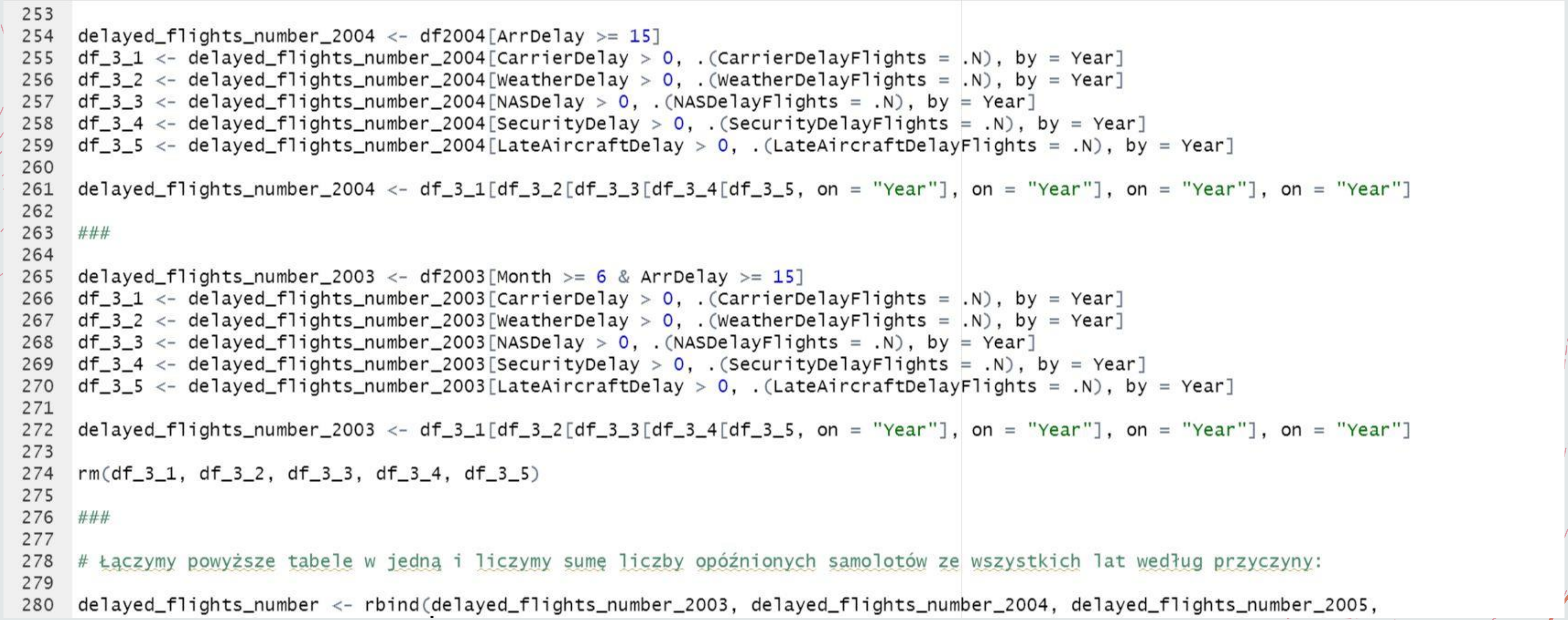
```r
plot_1


# ------------------------------------------------------------------------------------------- #
# Wykres 2.: Porównanie całkowitego czasu opóźnienia według przyczyny dla poszczególnych lat

delay_data <- rbind(sum_delay_by_cause_2003, sum_delay_by_cause_2004, sum_delay_by_cause_2005, sum_delay_by_cause_2006,
                    sum_delay_by_cause_2007, sum_delay_by_cause_2008)[, 1:6]
delay_data <- delay_data[, .(
  CarrierDelay = round(CarrierDelay / 60, 2) / 1000,
  WeatherDelay = round(WeatherDelay / 60, 2) / 1000,
  NASDelay = round(NASDelay / 60, 2) / 1000,
  SecurityDelay = round(SecurityDelay / 60, 2) / 1000,
  LateAircraftDelay = round(LateAircraftDelay / 60, 2) / 1000),
  by = Year]
delay_data <- as.data.frame(delay_data)

### wykres słupkowy: (próba utworzenia wykresu, w którym dla każdego roku jest 5 słupków odpowiadających całkowitemu czasowi
# opóźnienia dla każdej przyczyny)

# plot_2 <- ggplot(delay_data, aes(x = Year)) +
#  geom_bar(aes(y = CarrierDelay), stat = "identity", fill = "skyblue", width = 0.1) +
#  geom_bar(aes(y = WeatherDelay), stat = "identity", fill = "#FFCC99", width = 0.1) +
#  geom_bar(aes(y = NASDelay), stat = "identity", fill = "lightgreen", width = 0.1) +
#  geom_bar(aes(y = SecurityDelay), stat = "identity", fill = "orchid", width = 0.1) +
#  geom_bar(aes(y = LateAircraftDelay), stat = "identity", fill = "lightpink", width = 0.1) +
#  labs(x = "Year", y = "Delay Hours (1000 h)") +
#  scale_fill_manual(values = c("skyblue", "#FFCC99", "lightgreen", "orchid", "lightpink"),
```



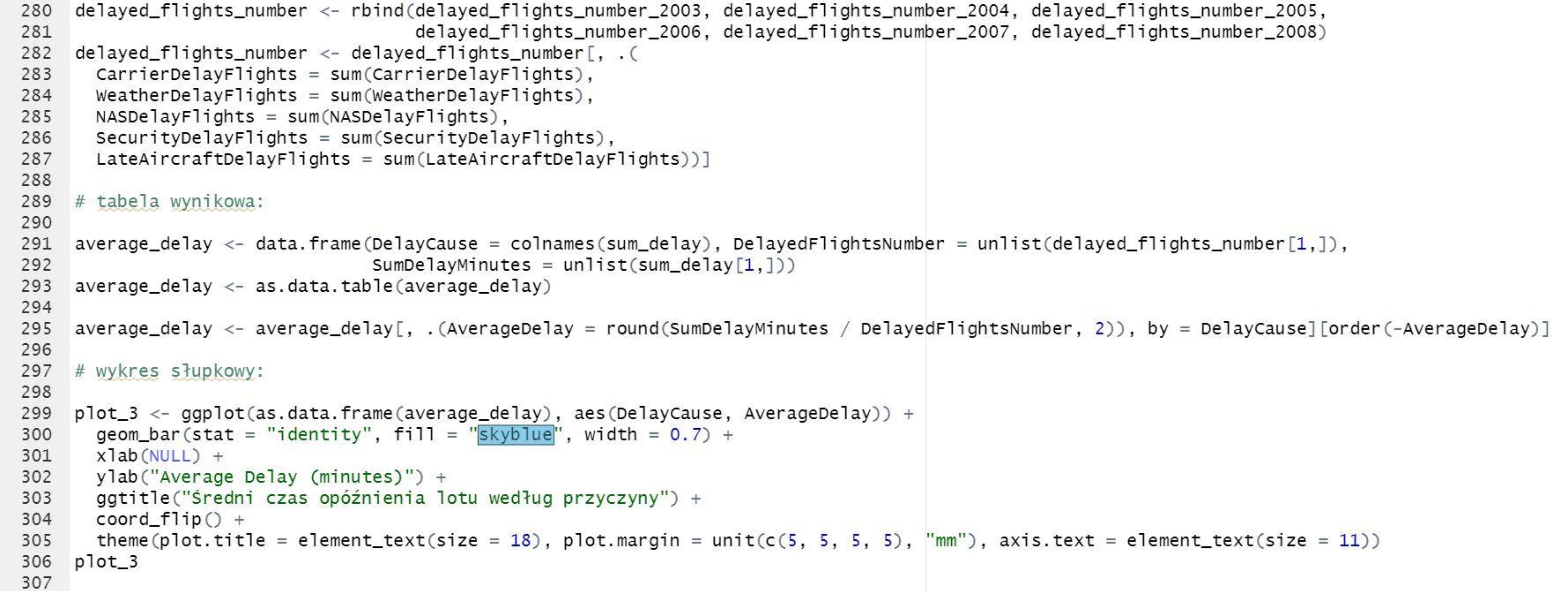
```
#                    labels = c("Carrier Delay", "Weather Delay", "NAS Delay", "Security Delay", "Late Aircraft Delay")) +
#  scale_y_continuous(limits = c(0, 700), breaks = seq(0, 700, by = 50)) +
#  labs(title = "Całkowite opóźnienie według przyczyny w latach 2003-2008") +
#  theme_minimal()
# plot_2


# ------------------------------------------------------------------- #
# Wykres 3.: Średni czas opóźnienia lotu dla poszczególnych przyczyn

# dla każdego roku: tabela z całkowitą liczbą opóźnionych samolotów według przyczyny

delayed_flights_number_2008 <- df2008[ArrDelay >= 15]
df_3_1 <- delayed_flights_number_2008[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
df_3_2 <- delayed_flights_number_2008[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
df_3_3 <- delayed_flights_number_2008[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
df_3_4 <- delayed_flights_number_2008[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
df_3_5 <- delayed_flights_number_2008[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]

delayed_flights_number_2008 <- df_3_1[df_3_2[df_3_3[df_3_4[df_3_5, on = "Year"], on = "Year"], on = "Year"], on = "Year"]

###

delayed_flights_number_2007 <- df2007[ArrDelay >= 15]
df_3_1 <- delayed_flights_number_2007[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
df_3_2 <- delayed_flights_number_2007[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
df_3_3 <- delayed_flights_number_2007[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
df_3_4 <- delayed_flights_number_2007[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
```



```
df_3_5 <- delayed_flights_number_2007[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]

delayed_flights_number_2007 <- df_3_1[df_3_2[df_3_3[df_3_4[df_3_5, on = "Year"], on = "Year"], on = "Year"], on = "Year"]

###

delayed_flights_number_2006 <- df2006[ArrDelay >= 15]
df_3_1 <- delayed_flights_number_2006[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
df_3_2 <- delayed_flights_number_2006[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
df_3_3 <- delayed_flights_number_2006[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
df_3_4 <- delayed_flights_number_2006[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
df_3_5 <- delayed_flights_number_2006[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]

delayed_flights_number_2006 <- df_3_1[df_3_2[df_3_3[df_3_4[df_3_5, on = "Year"], on = "Year"], on = "Year"], on = "Year"]

###

delayed_flights_number_2005 <- df2005[ArrDelay >= 15]
df_3_1 <- delayed_flights_number_2005[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
df_3_2 <- delayed_flights_number_2005[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
df_3_3 <- delayed_flights_number_2005[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
df_3_4 <- delayed_flights_number_2005[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
df_3_5 <- delayed_flights_number_2005[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]

delayed_flights_number_2005 <- df_3_1[df_3_2[df_3_3[df_3_4[df_3_5, on = "Year"], on = "Year"], on = "Year"], on = "Year"]

###
```

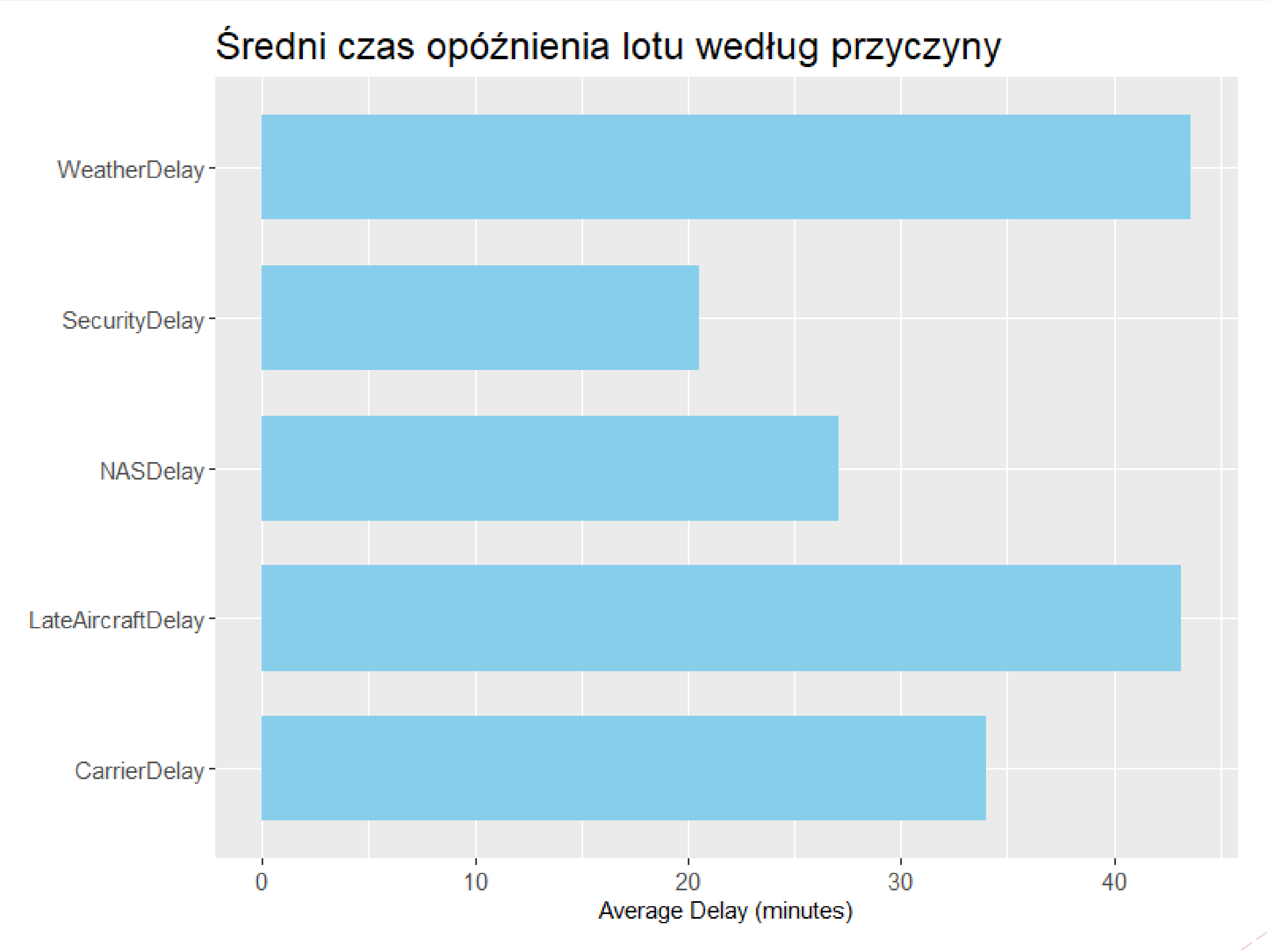


```
delayed_flights_number_2004 <- df2004[ArrDelay >= 15]
df_3_1 <- delayed_flights_number_2004[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
df_3_2 <- delayed_flights_number_2004[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
df_3_3 <- delayed_flights_number_2004[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
df_3_4 <- delayed_flights_number_2004[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
df_3_5 <- delayed_flights_number_2004[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]

delayed_flights_number_2004 <- df_3_1[df_3_2[df_3_3[df_3_4[df_3_5, on = "Year"], on = "Year"], on = "Year"], on = "Year"]

###

delayed_flights_number_2003 <- df2003[Month >= 6 & ArrDelay >= 15]
df_3_1 <- delayed_flights_number_2003[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
df_3_2 <- delayed_flights_number_2003[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
df_3_3 <- delayed_flights_number_2003[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
df_3_4 <- delayed_flights_number_2003[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
df_3_5 <- delayed_flights_number_2003[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]

delayed_flights_number_2003 <- df_3_1[df_3_2[df_3_3[df_3_4[df_3_5, on = "Year"], on = "Year"], on = "Year"], on = "Year"]

rm(df_3_1, df_3_2, df_3_3, df_3_4, df_3_5)

###

# łączymy powyższe tabele w jedną i liczymy sumę liczby opóźnionych samolotów ze wszystkich lat według przyczyny:

delayed_flights_number <- rbind(delayed_flights_number_2003, delayed_flights_number_2004, delayed_flights_number_2005,
```



```r
delayed_flights_number <- rbind(delayed_flights_number_2003, delayed_flights_number_2004, delayed_flights_number_2005,
                                delayed_flights_number_2006, delayed_flights_number_2007, delayed_flights_number_2008)
delayed_flights_number <- delayed_flights_number[, .(
  CarrierDelayFlights = sum(CarrierDelayFlights),
  WeatherDelayFlights = sum(WeatherDelayFlights),
  NASDelayFlights = sum(NASDelayFlights),
  SecurityDelayFlights = sum(SecurityDelayFlights),
  LateAircraftDelayFlights = sum(LateAircraftDelayFlights))]

# tabela wynikowa:

average_delay <- data.frame(DelayCause = colnames(sum_delay), DelayedFlightsNumber = unlist(delayed_flights_number[1,]),
                            SumDelayMinutes = unlist(sum_delay[1,]))
average_delay <- as.data.table(average_delay)

average_delay <- average_delay[, .(AverageDelay = round(SumDelayMinutes / DelayedFlightsNumber, 2)), by = DelayCause][order(-AverageDelay)]

# wykres słupkowy:

plot_3 <- ggplot(as.data.frame(average_delay), aes(DelayCause, AverageDelay)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  xlab(NULL) +
  ylab("Average Delay (minutes)") +
  ggtitle("Średni czas opóźnienia lotu według przyczyny") +
  coord_flip() +
  theme(plot.title = element_text(size = 18), plot.margin = unit(c(5, 5, 5, 5), "mm"), axis.text = element_text(size = 11))
plot_3

```

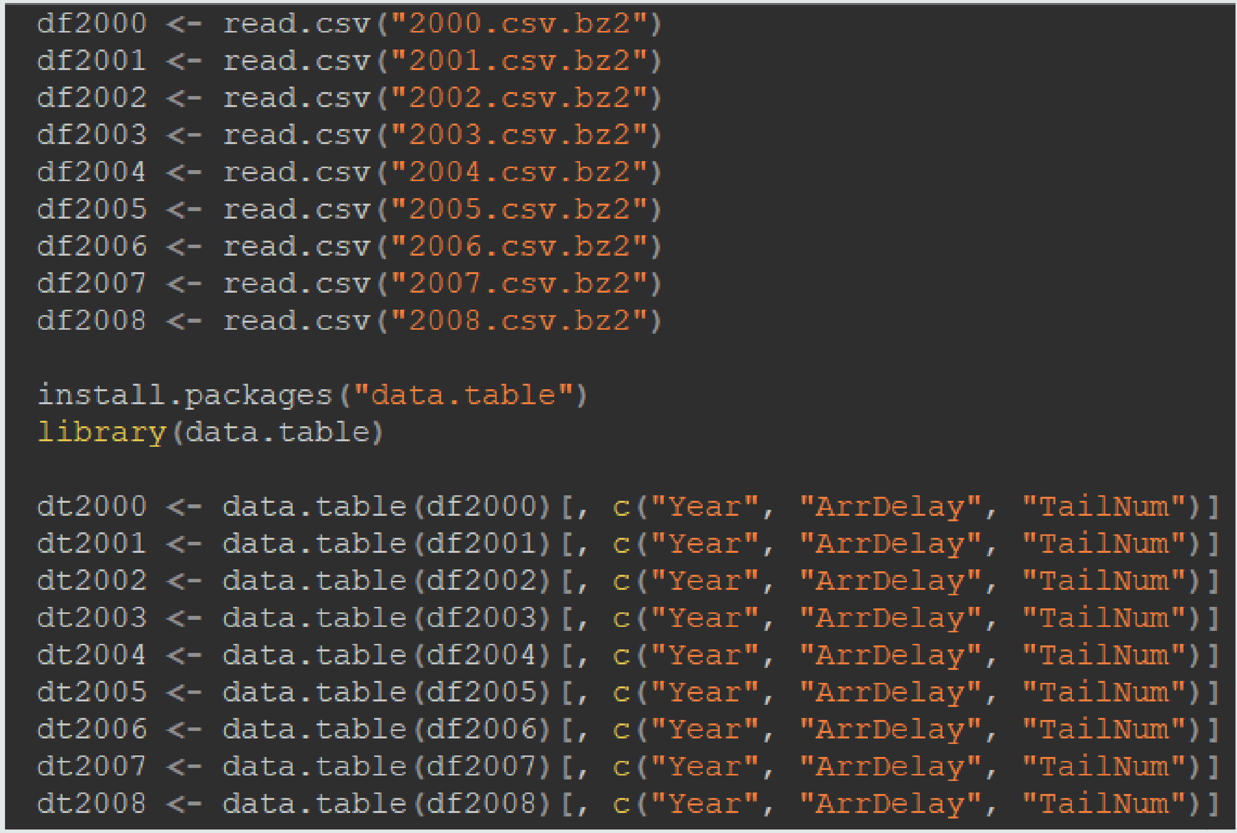
- Kategoria "WeatherDelay" nie pokazuje rzeczywistego wpływu pogody na opóźnienia.
  W ramach kategorii NAS istnieje osobna podkategoria dla pogody, która spowalnia działanie systemu, ale nie uniemożliwia lotów.



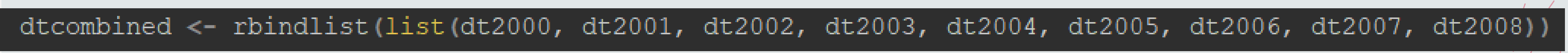
- Dane ze wszystkich lat 2003-2008

ANALIZA ZWIĄZKU MIĘDZY OPÓŹNIENIAMI A ROKIEM PRODUKCJI SAMOLOTU

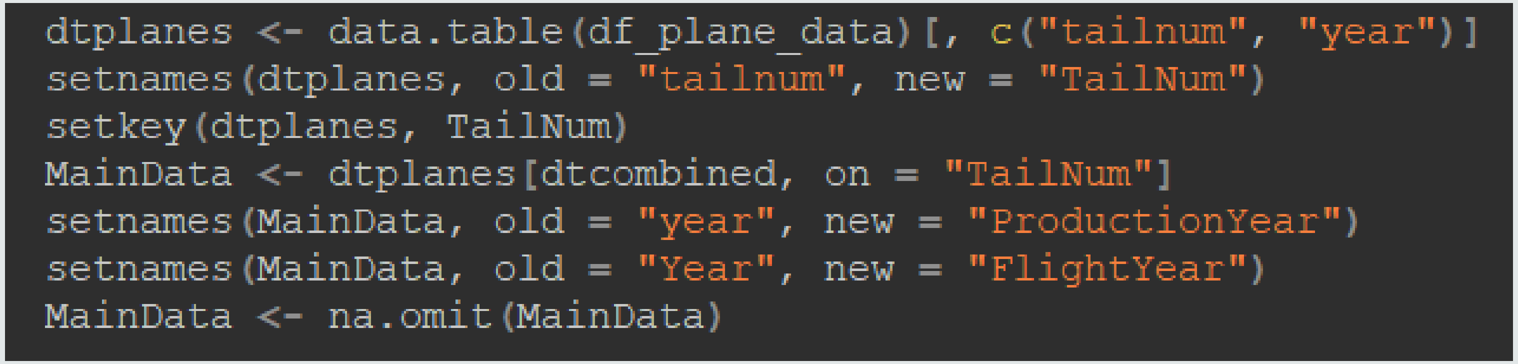
Obróbka danych



```
df2000 <- read.csv("2000.csv.bz2")
df2001 <- read.csv("2001.csv.bz2")
df2002 <- read.csv("2002.csv.bz2")
df2003 <- read.csv("2003.csv.bz2")
df2004 <- read.csv("2004.csv.bz2")
df2005 <- read.csv("2005.csv.bz2")
df2006 <- read.csv("2006.csv.bz2")
df2007 <- read.csv("2007.csv.bz2")
df2008 <- read.csv("2008.csv.bz2")

install.packages("data.table")
library(data.table)

dt2000 <- data.table(df2000)[, c("Year", "ArrDelay", "TailNum")]
dt2001 <- data.table(df2001)[, c("Year", "ArrDelay", "TailNum")]
dt2002 <- data.table(df2002)[, c("Year", "ArrDelay", "TailNum")]
dt2003 <- data.table(df2003)[, c("Year", "ArrDelay", "TailNum")]
dt2004 <- data.table(df2004)[, c("Year", "ArrDelay", "TailNum")]
dt2005 <- data.table(df2005)[, c("Year", "ArrDelay", "TailNum")]
dt2006 <- data.table(df2006)[, c("Year", "ArrDelay", "TailNum")]
dt2007 <- data.table(df2007)[, c("Year", "ArrDelay", "TailNum")]
dt2008 <- data.table(df2008)[, c("Year", "ArrDelay", "TailNum")]
```

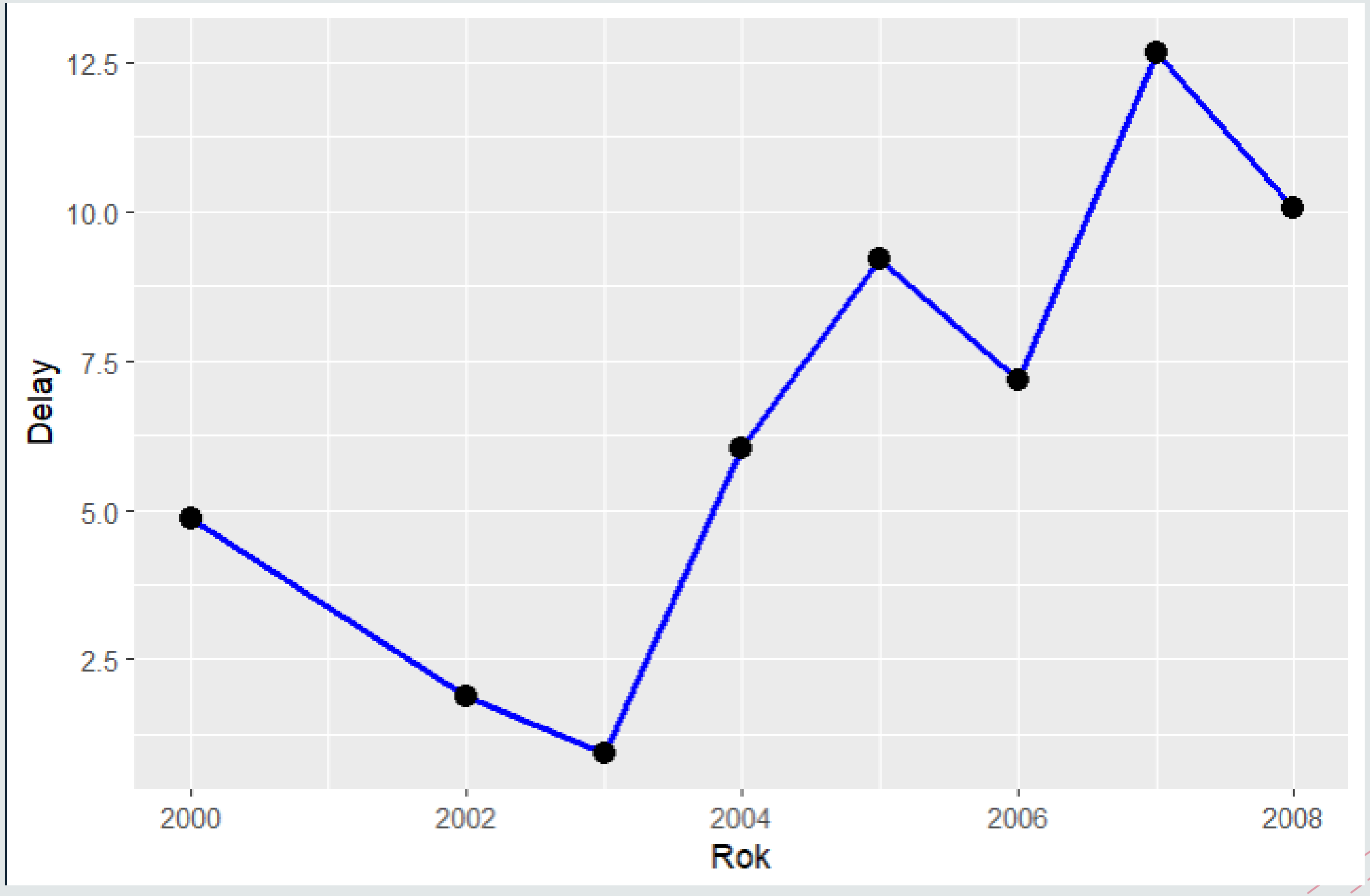


```
dtcombined <- rbindlist(list(dt2000, dt2001, dt2002, dt2003, dt2004, dt2005, dt2006, dt2007, dt2008))
```

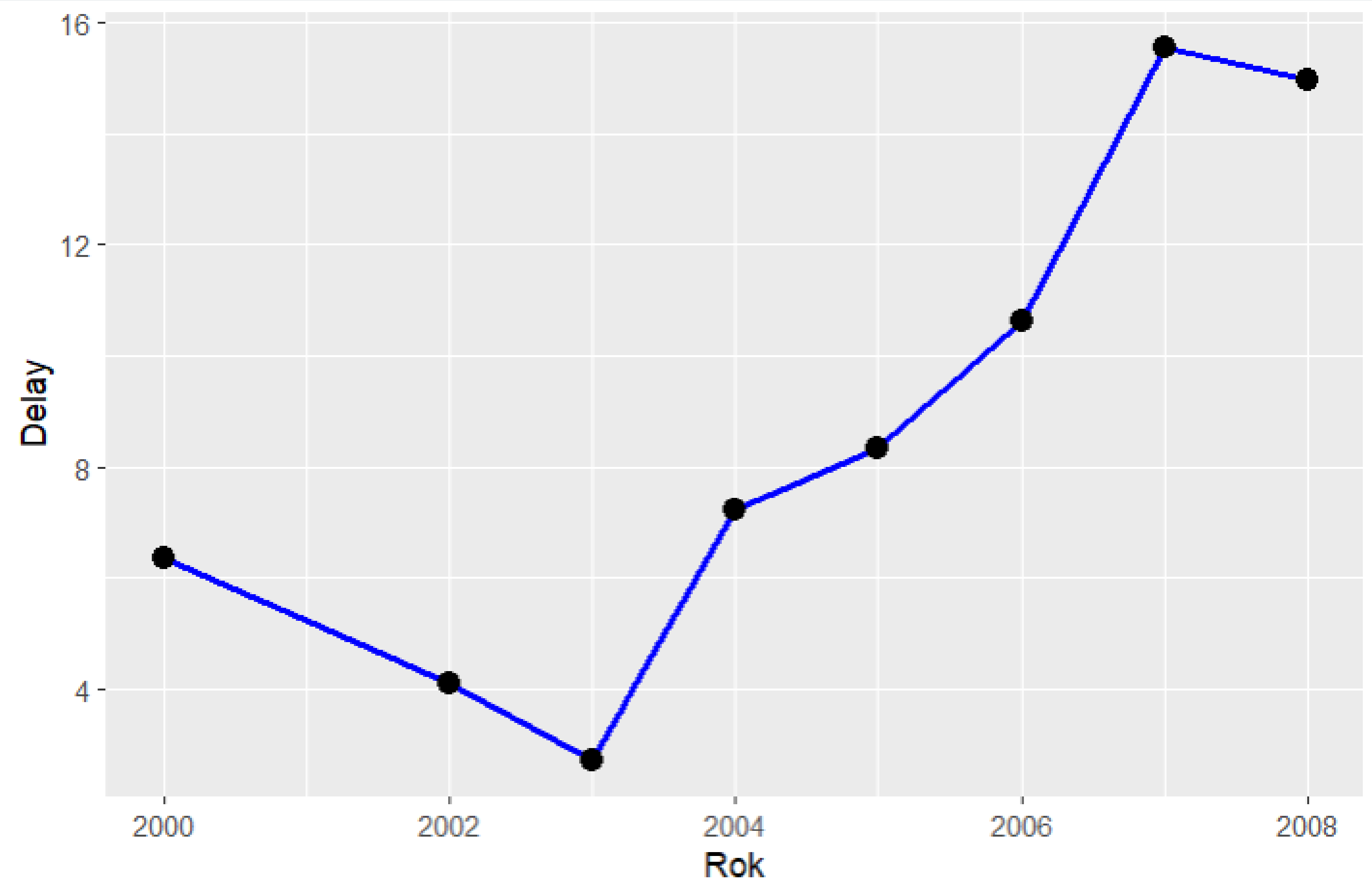


```
dtplanes <- data.table(df_plane_data)[, c("tailnum", "year")]
setnames(dtplanes, old = "tailnum", new = "TailNum")
setkey(dtplanes, TailNum)
MainData <- dtplanes[dtcombined, on = "TailNum"]
setnames(MainData, old = "year", new = "ProductionYear")
setnames(MainData, old = "Year", new = "FlightYear")
MainData <- na.omit(MainData)
```

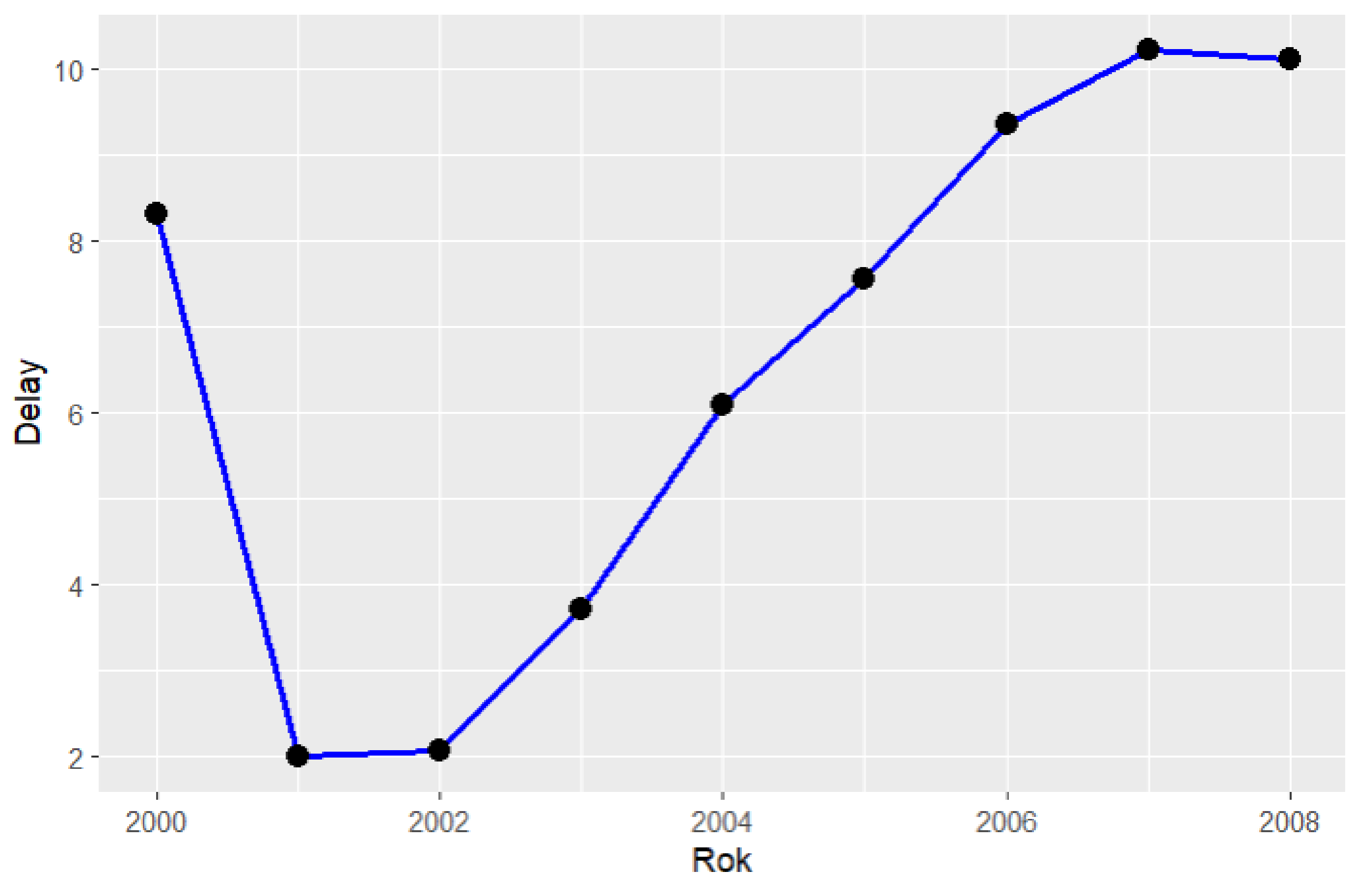
Wykres średniego opóźnienia dla samolotów wyprodukowanych w 1970



Wykres średniego opóźnienia dla samolotów wyprodukowanych w 1980



Wykres średniego opóźnienia dla samolotów wyprodukowanych w 1999


Delay
Rok