

# Projekt grupowy PDU

Julia Dollani, Aleksandra Wójcik, Jan Opala



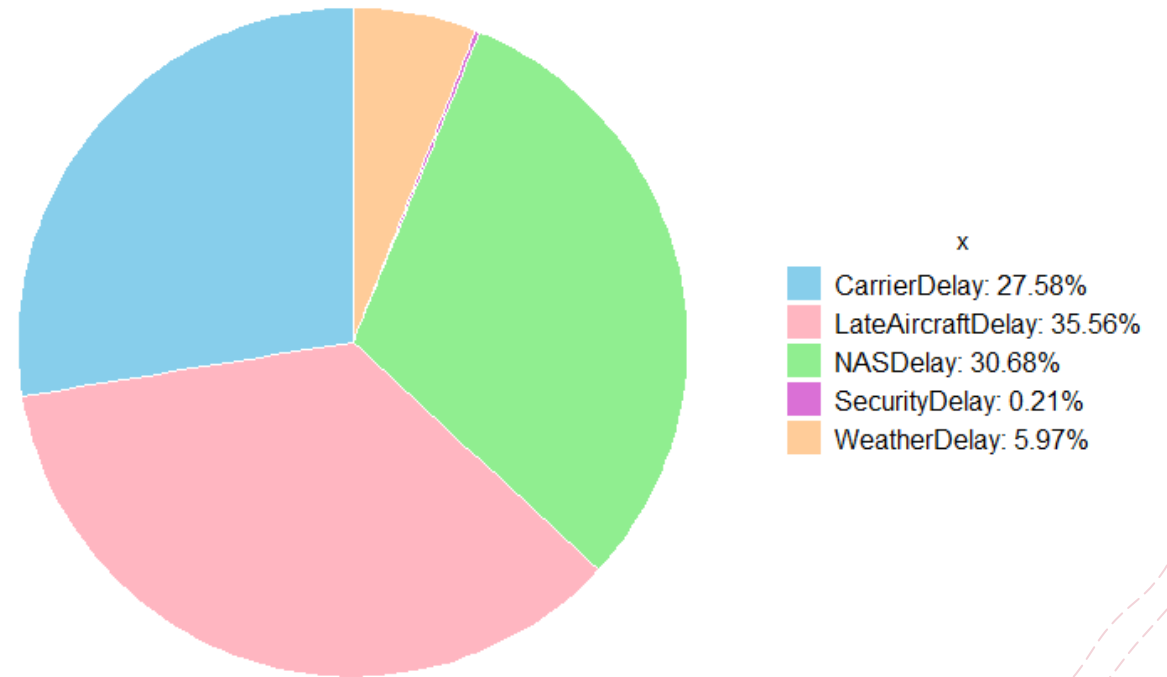
# ANALIZA WPŁYWU RÓŻNYCH CZYNNIKÓW NA OPÓŹNIENIA LOTÓW

### 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



```
1 # install.packages("tidyverse")
2
3 library("data.table")
4 library("dplyr")
5 library(ggplot2)
6
7 setwd("C:\\Users\\Ola\\Desktop\\Studia PW\\PDU")
8
9 Airports <- as.data.table(read.csv("airports.csv"))
10 Carriers <- as.data.table(read.csv("carriers.csv"))
11 # Plane_data <- as.data.table(read.csv("plane-data.csv"))
12 Variable_descriptions <- as.data.table(read.csv("variable-descriptions.csv"))
13 df2008 <- as.data.table(read.csv("2008.csv.bz2"))
14 df2007 <- as.data.table(read.csv("2007.csv.bz2"))
15 df2006 <- as.data.table(read.csv("2006.csv.bz2"))
16 df2005 <- as.data.table(read.csv("2005.csv.bz2"))
17 df2004 <- as.data.table(read.csv("2004.csv.bz2"))
18 df2003 <- as.data.table(read.csv("2003.csv.bz2"))
19
20
21 # ----- #
22 # wykres 1.: Procentowy udział przyczyn opóźnienia lotów w całkowitym czasie opóźnienia dla lat 2003-2008
23 # (dane o przyczynie opóźnienia są zbierane od czerwca 2003 roku)
24
25 # dla każdego roku: tabela z całkowitym czasem opóźnienia według przyczyny i kolumna z sumą wszystkich minut opóźnienia w danym roku
26
27 sum_delay_by_cause_2008 <- df2008[ArrDelay >= 15]
28
```



```
28
29 sum_delay_by_cause_2008 <- sum_delay_by_cause_2008[, .(
30   CarrierDelay = sum(CarrierDelay),
31   WeatherDelay = sum(WeatherDelay),
32   NASDelay = sum(NASDelay),
33   SecurityDelay = sum(SecurityDelay),
34   LateAircraftDelay = sum(LateAircraftDelay)),
35   by = Year]
36
37 total_delay <- sum(sum_delay_by_cause_2008[, 2:6])
38
39 sum_delay_by_cause_2008 <- mutate(sum_delay_by_cause_2008, TotalDelay = total_delay)
40
41 rm(total_delay)
42
43 ###
44
45 sum_delay_by_cause_2007 <- df2007[ArrDelay >= 15]
46
47 sum_delay_by_cause_2007 <- sum_delay_by_cause_2007[, .(
48   CarrierDelay = sum(CarrierDelay),
49   WeatherDelay = sum(WeatherDelay),
50   NASDelay = sum(NASDelay),
51   SecurityDelay = sum(SecurityDelay),
52   LateAircraftDelay = sum(LateAircraftDelay)),
53   by = Year]
54
55 total_delay <- sum(sum_delay_by_cause_2007[, 2:6])
56
```

```
56
57 sum_delay_by_cause_2007 <- mutate(sum_delay_by_cause_2007, TotalDelay = total_delay)
58
59 rm(total_delay)
60
61 ###
62
63 sum_delay_by_cause_2006 <- df2006[ArrDelay >= 15]
64
65 sum_delay_by_cause_2006 <- sum_delay_by_cause_2006[, .(
66   CarrierDelay = sum(CarrierDelay),
67   WeatherDelay = sum(WeatherDelay),
68   NASDelay = sum(NASDelay),
69   SecurityDelay = sum(SecurityDelay),
70   LateAircraftDelay = sum(LateAircraftDelay)),
71   by = Year]
72
73 total_delay <- sum(sum_delay_by_cause_2006[, 2:6])
74
75 sum_delay_by_cause_2006 <- mutate(sum_delay_by_cause_2006, TotalDelay = total_delay)
76
77 rm(total_delay)
78
79 ###
80
81 sum_delay_by_cause_2005 <- df2005[ArrDelay >= 15]
82
83 sum_delay_by_cause_2005 <- sum_delay_by_cause_2005[, .(
84   CarrierDelay = sum(CarrierDelay),
```

```
85   WeatherDelay = sum(WeatherDelay),
86   NASDelay = sum(NASDelay),
87   SecurityDelay = sum(SecurityDelay),
88   LateAircraftDelay = sum(LateAircraftDelay)),
89   by = Year]
90
91 total_delay <- sum(sum_delay_by_cause_2005[, 2:6])
92
93 sum_delay_by_cause_2005 <- mutate(sum_delay_by_cause_2005, TotalDelay = total_delay)
94
95 rm(total_delay)
96
97 ###
98
99 sum_delay_by_cause_2004 <- df2004[ArrDelay >= 15]
100
101 sum_delay_by_cause_2004 <- sum_delay_by_cause_2004[, .(
102   CarrierDelay = sum(CarrierDelay),
103   WeatherDelay = sum(WeatherDelay),
104   NASDelay = sum(NASDelay),
105   SecurityDelay = sum(SecurityDelay),
106   LateAircraftDelay = sum(LateAircraftDelay)),
107   by = Year]
108
109 total_delay <- sum(sum_delay_by_cause_2004[, 2:6])
110
111 sum_delay_by_cause_2004 <- mutate(sum_delay_by_cause_2004, TotalDelay = total_delay)
112
```

```
113 rm(total_delay)
114
115 ###
116
117 sum_delay_by_cause_2003 <- df2003[Month >= 6 & ArrDelay >= 15]
118
119 sum_delay_by_cause_2003 <- sum_delay_by_cause_2003[, .(
120   CarrierDelay = sum(CarrierDelay),
121   WeatherDelay = sum(WeatherDelay),
122   NASDelay = sum(NASDelay),
123   SecurityDelay = sum(SecurityDelay),
124   LateAircraftDelay = sum(LateAircraftDelay)),
125   by = Year]
126
127 total_delay <- sum(sum_delay_by_cause_2003[, 2:6])
128
129 sum_delay_by_cause_2003 <- mutate(sum_delay_by_cause_2003, TotalDelay = total_delay)
130
131 rm(total_delay)
132
133 ### tabela wynikowa:
134
135 sum_delay <- rbind(sum_delay_by_cause_2003, sum_delay_by_cause_2004, sum_delay_by_cause_2005, sum_delay_by_cause_2006,
136   sum_delay_by_cause_2007, sum_delay_by_cause_2008)
137
138 total_delay <- sum(sum_delay[, TotalDelay])
139
140 sum_delay <- sum_delay[, .(
```



```

141   CarrierDelay = sum(CarrierDelay),
142   WeatherDelay = sum(WeatherDelay),
143   NASDelay = sum(NASDelay),
144   SecurityDelay = sum(SecurityDelay),
145   LateAircraftDelay = sum(LateAircraftDelay))]]
146
147   agg_result_1 <- sum_delay[, .(
148     CarrierDelay = (CarrierDelay / total_delay) * 100,
149     WeatherDelay = (WeatherDelay / total_delay) * 100,
150     NASDelay = (NASDelay / total_delay) * 100,
151     SecurityDelay = (SecurityDelay / total_delay) * 100,
152     LateAircraftDelay = (LateAircraftDelay / total_delay) * 100)]
153
154   rm(total_delay)
155
156   ### wykres kołowy:
157
158   values <- unlist(agg_result_1[1,])
159   labels_2 <- paste(round(values, 2), "%", sep = "")
160   labels_1 <- paste(paste(names(agg_result_1), ": ", sep = ""), labels_2, sep = "")
161
162   plot_1 <- ggplot(data.frame(x = labels_1, y = values), aes(x = "", y = y, fill = x)) +
163     geom_bar(width = 1, stat = "identity", color = "white") +
164     coord_polar(theta = "y") +
165     scale_fill_manual(values = c("skyblue", "lightpink", "lightgreen", "orchid", "#FFCC99")) +
166     labs(title = "Procentowy udział przyczyn opóźnień lotów \nw całkowitym czasie opóźnień w latach 2003-2008") +
167     theme_void() +
168     theme(plot.title = element_text(size = 20), plot.margin = unit(c(5, 5, 5, 5), "mm"), legend.text = element_text(size = 12),
169           legend.title = NULL)

```

```
170 plot_1
171
172
173 # ----- #
174 # wykres 2.: Porównanie całkowitego czasu opóźnienia według przyczyny dla poszczególnych lat
175
176 delay_data <- rbind(sum_delay_by_cause_2003, sum_delay_by_cause_2004, sum_delay_by_cause_2005, sum_delay_by_cause_2006,
177                    sum_delay_by_cause_2007, sum_delay_by_cause_2008)[, 1:6]
178 delay_data <- delay_data[, .(
179   CarrierDelay = round(CarrierDelay / 60, 2) / 1000,
180   WeatherDelay = round(WeatherDelay / 60, 2) / 1000,
181   NASDelay = round(NASDelay / 60, 2) / 1000,
182   SecurityDelay = round(SecurityDelay / 60, 2) / 1000,
183   LateAircraftDelay = round(LateAircraftDelay / 60, 2) / 1000),
184   by = Year]
185 delay_data <- as.data.frame(delay_data)
186
187 ### wykres słupkowy: (próba utworzenia wykresu, w którym dla każdego roku jest 5 słupków odpowiadających całkowitemu czasowi
188 # opóźnienia dla każdej przyczyny)
189
190 # plot_2 <- ggplot(delay_data, aes(x = Year)) +
191 #   geom_bar(aes(y = CarrierDelay), stat = "identity", fill = "skyblue", width = 0.1) +
192 #   geom_bar(aes(y = WeatherDelay), stat = "identity", fill = "#FFCC99", width = 0.1) +
193 #   geom_bar(aes(y = NASDelay), stat = "identity", fill = "lightgreen", width = 0.1) +
194 #   geom_bar(aes(y = SecurityDelay), stat = "identity", fill = "orchid", width = 0.1) +
195 #   geom_bar(aes(y = LateAircraftDelay), stat = "identity", fill = "lightpink", width = 0.1) +
196 #   labs(x = "Year", y = "Delay Hours (1000 h)") +
197 #   scale_fill_manual(values = c("skyblue", "#FFCC99", "lightgreen", "orchid", "lightpink"),
```



```

198 # labels = c("Carrier Delay", "Weather Delay", "NAS Delay", "Security Delay", "Late Aircraft Delay")) +
199 # scale_y_continuous(limits = c(0, 700), breaks = seq(0, 700, by = 50)) +
200 # labs(title = "Całkowite opóźnienie według przyczyny w latach 2003-2008") +
201 # theme_minimal()
202 # plot_2
203
204
205 # ----- #
206 # wykres 3.: Średni czas opóźnienia lotu dla poszczególnych przyczyn
207
208 # dla każdego roku: tabela z całkowitą liczbą opóźnionych samolotów według przyczyny
209
210 delayed_flights_number_2008 <- df2008[ArrDelay >= 15]
211 df_3_1 <- delayed_flights_number_2008[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
212 df_3_2 <- delayed_flights_number_2008[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
213 df_3_3 <- delayed_flights_number_2008[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
214 df_3_4 <- delayed_flights_number_2008[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
215 df_3_5 <- delayed_flights_number_2008[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]
216
217 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"]
218
219 ###
220
221 delayed_flights_number_2007 <- df2007[ArrDelay >= 15]
222 df_3_1 <- delayed_flights_number_2007[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
223 df_3_2 <- delayed_flights_number_2007[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
224 df_3_3 <- delayed_flights_number_2007[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
225 df_3_4 <- delayed_flights_number_2007[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]

```

```
226 df_3_5 <- delayed_flights_number_2007[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]
227
228 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"]
229
230 ###
231
232 delayed_flights_number_2006 <- df2006[ArrDelay >= 15]
233 df_3_1 <- delayed_flights_number_2006[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
234 df_3_2 <- delayed_flights_number_2006[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
235 df_3_3 <- delayed_flights_number_2006[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
236 df_3_4 <- delayed_flights_number_2006[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
237 df_3_5 <- delayed_flights_number_2006[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]
238
239 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"]
240
241 ###
242
243 delayed_flights_number_2005 <- df2005[ArrDelay >= 15]
244 df_3_1 <- delayed_flights_number_2005[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
245 df_3_2 <- delayed_flights_number_2005[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
246 df_3_3 <- delayed_flights_number_2005[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
247 df_3_4 <- delayed_flights_number_2005[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
248 df_3_5 <- delayed_flights_number_2005[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]
249
250 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"]
251
252 ###
```



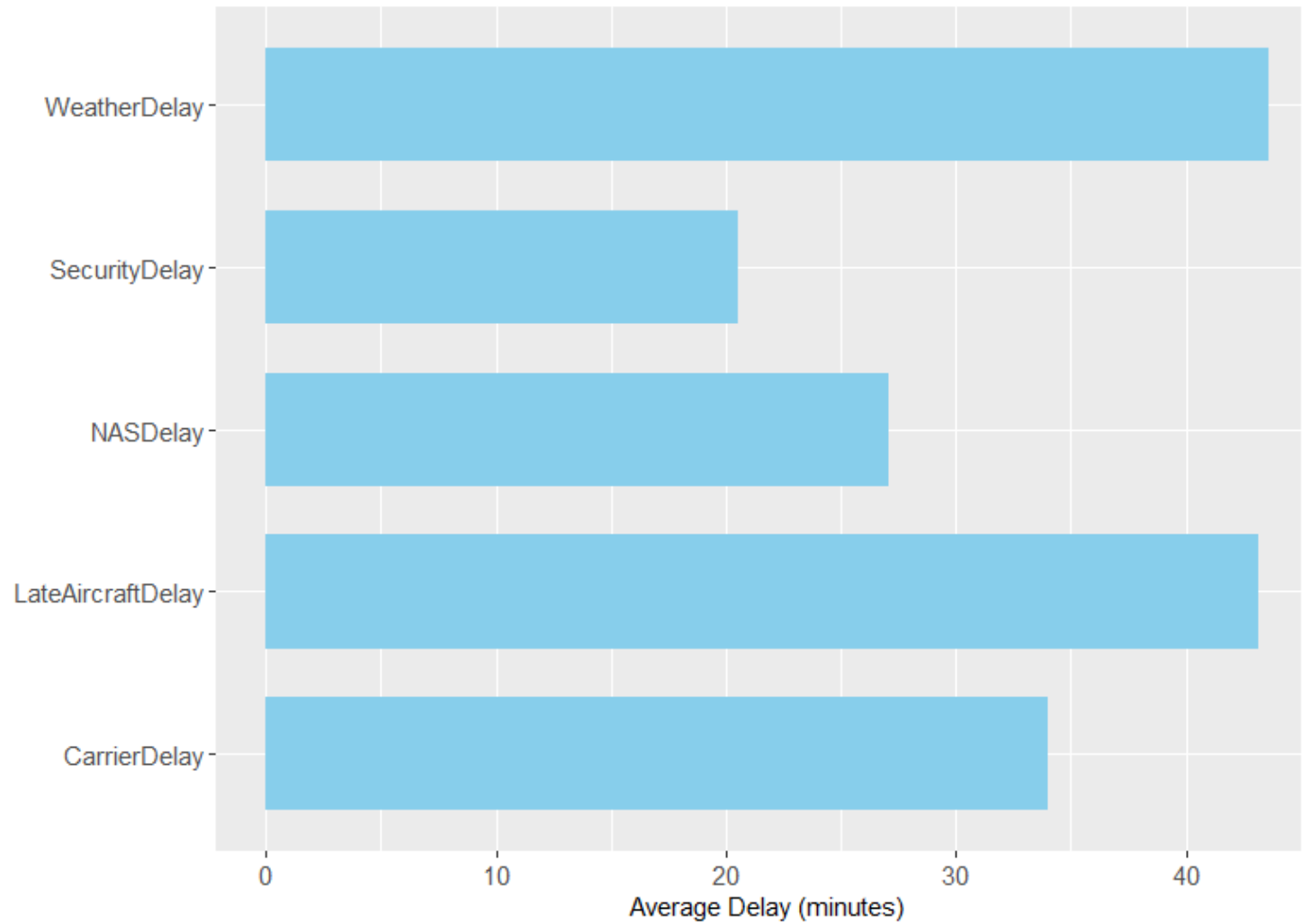
```
253
254 delayed_flights_number_2004 <- df2004[ArrDelay >= 15]
255 df_3_1 <- delayed_flights_number_2004[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
256 df_3_2 <- delayed_flights_number_2004[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
257 df_3_3 <- delayed_flights_number_2004[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
258 df_3_4 <- delayed_flights_number_2004[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
259 df_3_5 <- delayed_flights_number_2004[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]
260
261 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"]
262
263 ###
264
265 delayed_flights_number_2003 <- df2003[Month >= 6 & ArrDelay >= 15]
266 df_3_1 <- delayed_flights_number_2003[CarrierDelay > 0, .(CarrierDelayFlights = .N), by = Year]
267 df_3_2 <- delayed_flights_number_2003[WeatherDelay > 0, .(WeatherDelayFlights = .N), by = Year]
268 df_3_3 <- delayed_flights_number_2003[NASDelay > 0, .(NASDelayFlights = .N), by = Year]
269 df_3_4 <- delayed_flights_number_2003[SecurityDelay > 0, .(SecurityDelayFlights = .N), by = Year]
270 df_3_5 <- delayed_flights_number_2003[LateAircraftDelay > 0, .(LateAircraftDelayFlights = .N), by = Year]
271
272 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"]
273
274 rm(df_3_1, df_3_2, df_3_3, df_3_4, df_3_5)
275
276 ###
277
278 # łączymy powyższe tabele w jedną i liczymy sumę liczby opóźnionych samolotów ze wszystkich lat według przyczyny:
279
280 delayed_flights_number <- rbind(delayed_flights_number_2003, delayed_flights_number_2004, delayed_flights_number_2005,
```

```
280 delayed_flights_number <- rbind(delayed_flights_number_2003, delayed_flights_number_2004, delayed_flights_number_2005,
281                                 delayed_flights_number_2006, delayed_flights_number_2007, delayed_flights_number_2008)
282 delayed_flights_number <- delayed_flights_number[, .(
283   CarrierDelayFlights = sum(CarrierDelayFlights),
284   WeatherDelayFlights = sum(WeatherDelayFlights),
285   NASDelayFlights = sum(NASDelayFlights),
286   SecurityDelayFlights = sum(SecurityDelayFlights),
287   LateAircraftDelayFlights = sum(LateAircraftDelayFlights))]
288
289 # tabela wynikowa:
290
291 average_delay <- data.frame(DelayCause = colnames(sum_delay), DelayedFlightsNumber = unlist(delayed_flights_number[1,]),
292                             SumDelayMinutes = unlist(sum_delay[1,]))
293 average_delay <- as.data.table(average_delay)
294
295 average_delay <- average_delay[, .(AverageDelay = round(SumDelayMinutes / DelayedFlightsNumber, 2)), by = DelayCause][order(-AverageDelay)]
296
297 # wykres słupkowy:
298
299 plot_3 <- ggplot(as.data.frame(average_delay), aes(DelayCause, AverageDelay)) +
300   geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
301   xlab(NULL) +
302   ylab("Average Delay (minutes)") +
303   ggtitle("Średni czas opóźnienia lotu według przyczyny") +
304   coord_flip() +
305   theme(plot.title = element_text(size = 18), plot.margin = unit(c(5, 5, 5, 5), "mm"), axis.text = element_text(size = 11))
306 plot_3
307
```

- + 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.



### Średni czas opóźnienia lotu według przyczyny



+ Dane ze wszystkich lat 2003-2008





# ANALIZA ZWIĄZKU MIĘDZY OPÓŹNIENIAMI A ROKIEM PRODUKCJI SAMOŁOTU

# 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")
```

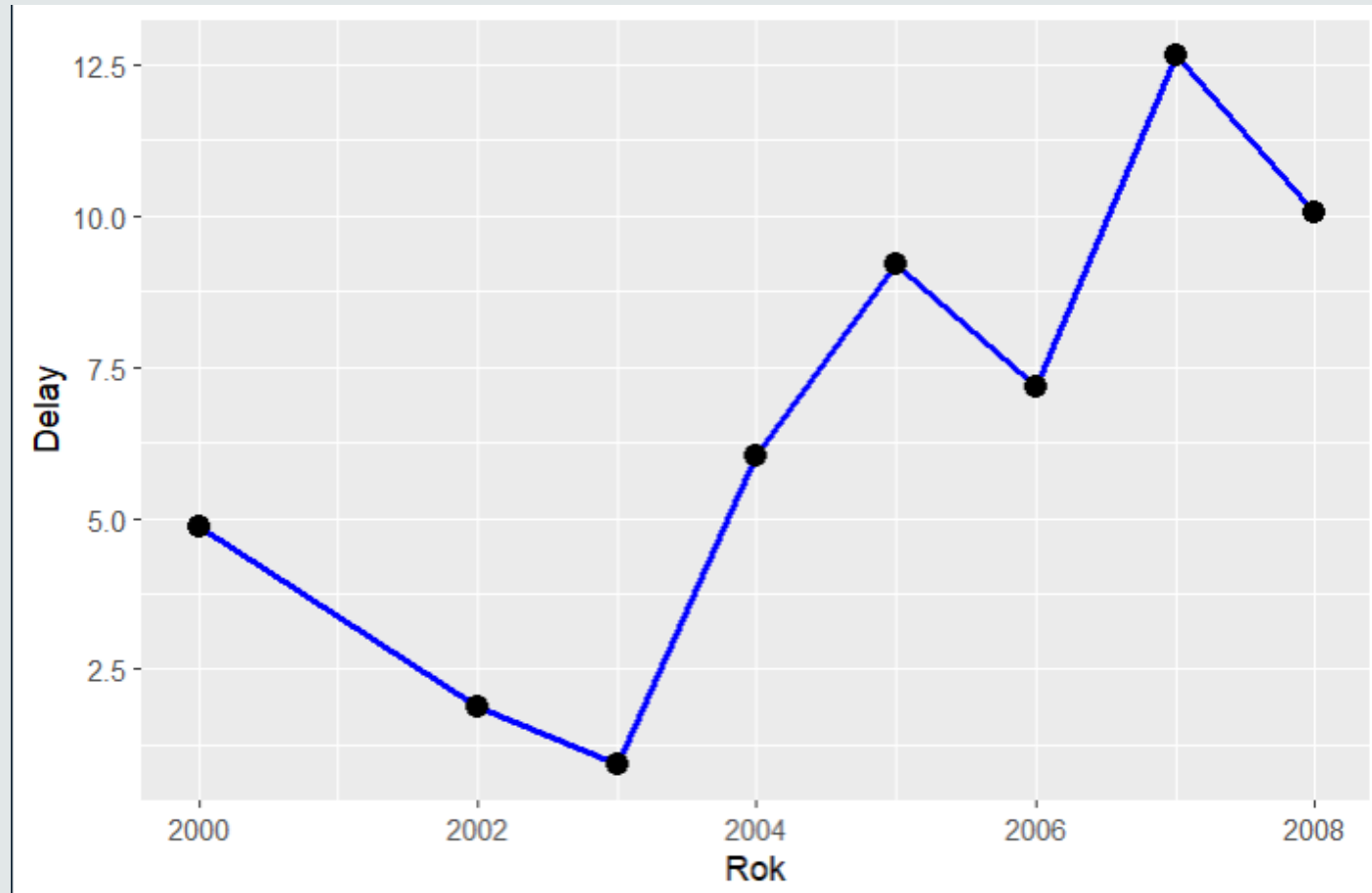
```
install.packages("data.table")
library(data.table)
```

```
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")]
```

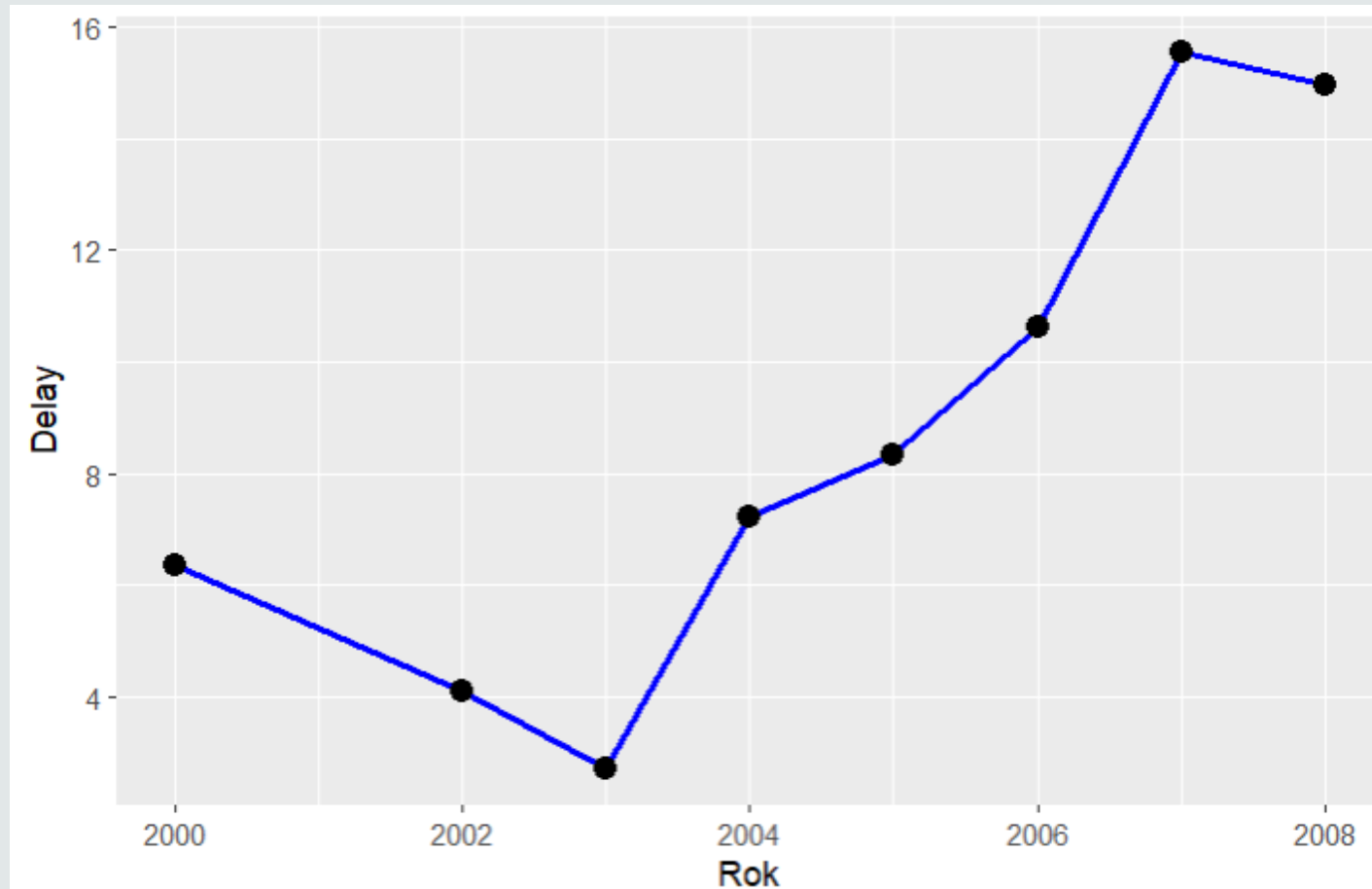
```
dtcombined <- rbindlist(list(dt2000, dt2001, dt2002, dt2003, dt2004, dt2005, dt2006, dt2007, dt2008))
```

```
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

