Recidivism Survival Intelligence

Kaplan-Meier, Log-Rank and Cox PH Modeling (Portfolio Report)

Executive Snapshot

- **Goal:** Quantify time-to-event dynamics and subgroup differences in re-arrest risk.
- **Methods:** Descriptive EDA, Kaplan-Meier curves, log-rank tests, Cox proportional hazards model.
- **Deliverables:** Reproducible R workflow, saved plots/tables, hazard ratios and PH diagnostics.

What's inside

- **1) EDA:** Distributions and group shares for key categorical and count predictors.
- **2) Survival Curves:** Kaplan-Meier survival curves across covariates with risk tables.
- **3) Cox Model:** Multivariate hazard ratios with confidence intervals for interpretation.
- **4) Diagnostics:** Proportional hazards checks and exported model summary tables.

Survival Analysis Portfolio (T4)

Time to Arrest — Kaplan–Meier and Cox Proportional Hazards

Setup & Data Preparation

```
suppressPackageStartupMessages({
  library(survival)
  library(survminer)
  library(dplyr)
  library(ggplot2)
  library(scales)
  library(broom)
})

out <- "outputs_T4"
if (!dir.exists(out)) dir.create(out, recursive = TRUE)

th <- function() {
  theme_minimal(base_size = 12) +
    theme(
      plot.title = element_text(face = "bold"),
      legend.position = "bottom",
      panel.grid.minor = element_blank()
    )
}

savep <- function(p, name, w=8, h=5, dpi=220) {
  ggsave(file.path(out, name), p, width=w, height=h, dpi=dpi)
  p
}

d <- read.table("Rossi.txt", header = TRUE)

d <- d %>%
  mutate(
    fin  = factor(fin,  levels=c(0,1), labels=c("No","Yes")),
    wexp = factor(wexp, levels=c(0,1), labels=c("No","Yes")),
    mar  = factor(mar,  levels=c(0,1), labels=c("No","Yes")),
    paro = factor(paro, levels=c(0,1), labels=c("No","Yes")),
    race = factor(race),
    educ = factor(educ),
    week = as.numeric(week),
    arrest = as.integer(arrest)
  )

stopifnot(all(d$week > 0), all(d$arrest %in% c(0,1)))
```

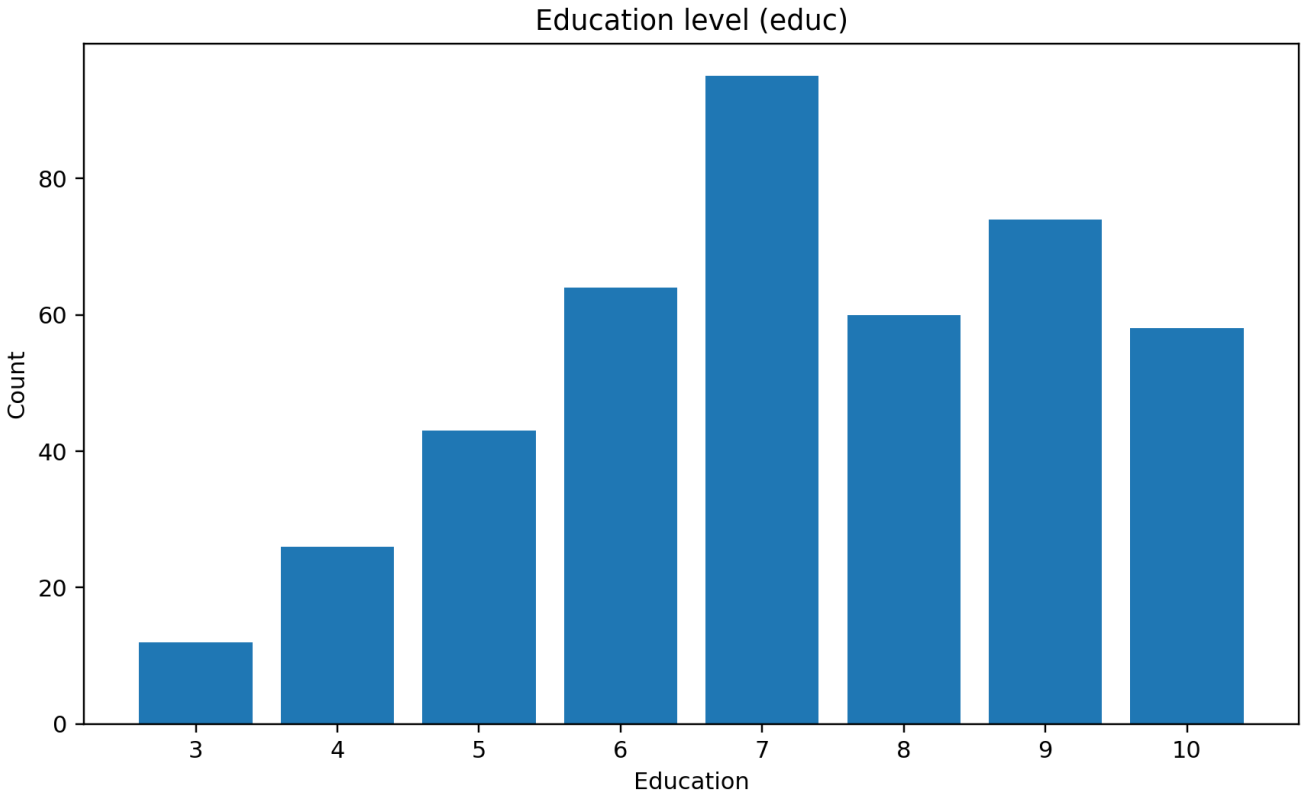
Interpretation: The dataset is validated (positive follow-up time, binary event) and categorical predictors are standardized for consistent reporting.

Helper: Distribution Table

```
share <- function(x) {
  tibble(v=x) %>%
    count(v) %>%
    mutate(p = n/sum(n)) %>%
    arrange(desc(n))
}
```

Education Distribution (educ)

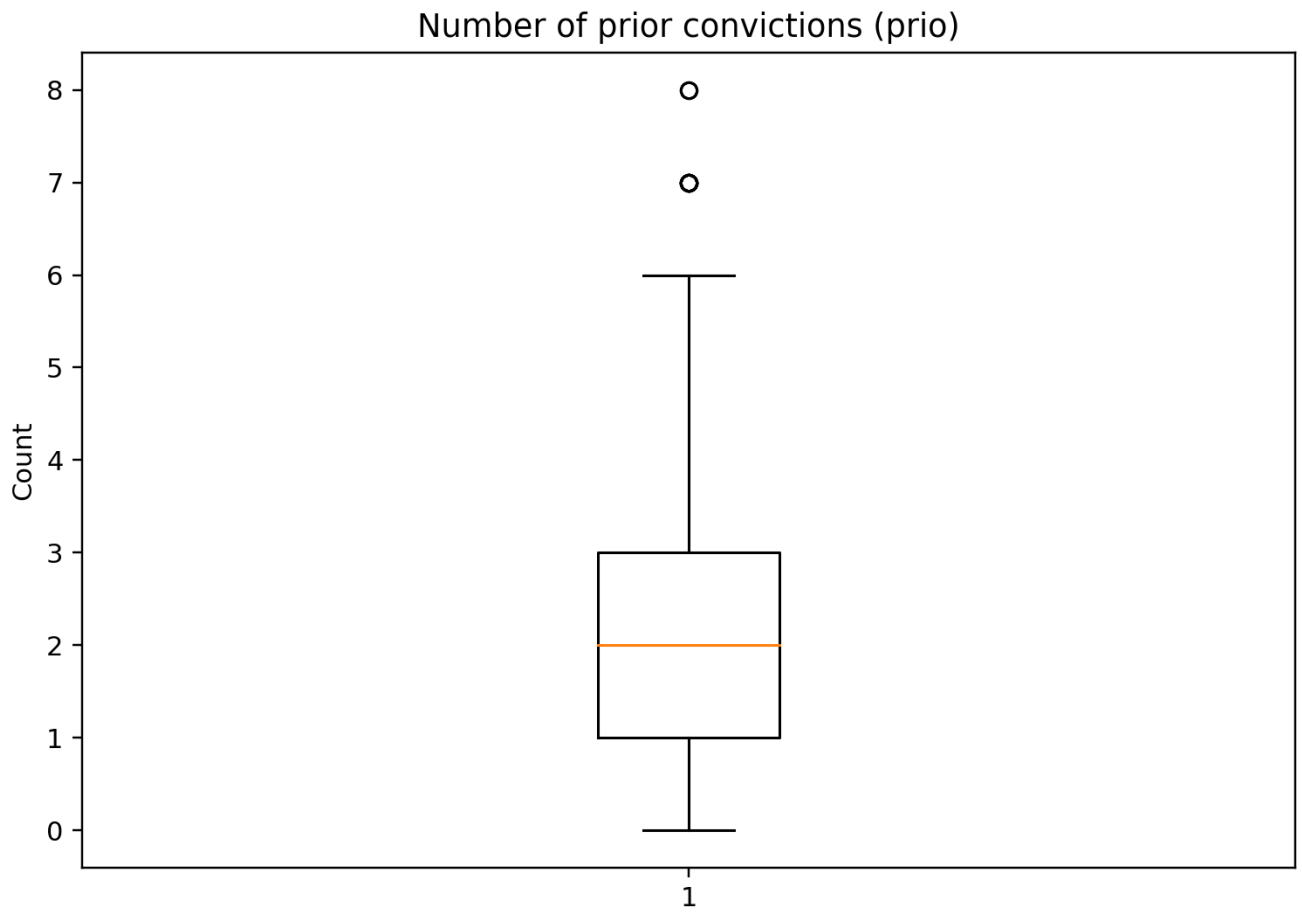
```
ed <- share(d$educ) %>% mutate(ed = as.numeric(as.character(v)))

p_educ <- ggplot(ed, aes(ed, p)) +
  geom_col(alpha=.9) +
  scale_x_continuous(breaks=sort(unique(ed$ed))) +
  scale_y_continuous(labels=percent_format(1),
                     expand=expansion(mult=c(0, .10))) +
  labs(title="educ", x="educ", y="Share") +
  th()

savep(p_educ, "06_educ.png", w=8, h=4.5)
print(p_educ)
```



Interpretation: Education shows clear concentration around mid-to-high levels, providing enough variability to support group-based survival comparisons.

Prior Convictions (prio)

```
p_prio <- ggplot(d, aes(x="", y=prio)) +
  geom_boxplot(outlier.alpha=.35) +
  labs(title="prio", x=NULL, y="Count") +
  th()

savep(p_prio, "07_prio.png", w=6.5, h=4.5)
print(p_prio)
```



Interpretation: Prior convictions exhibit right-skew and outliers, suggesting heterogeneous baseline risk that is informative for hazard modeling.

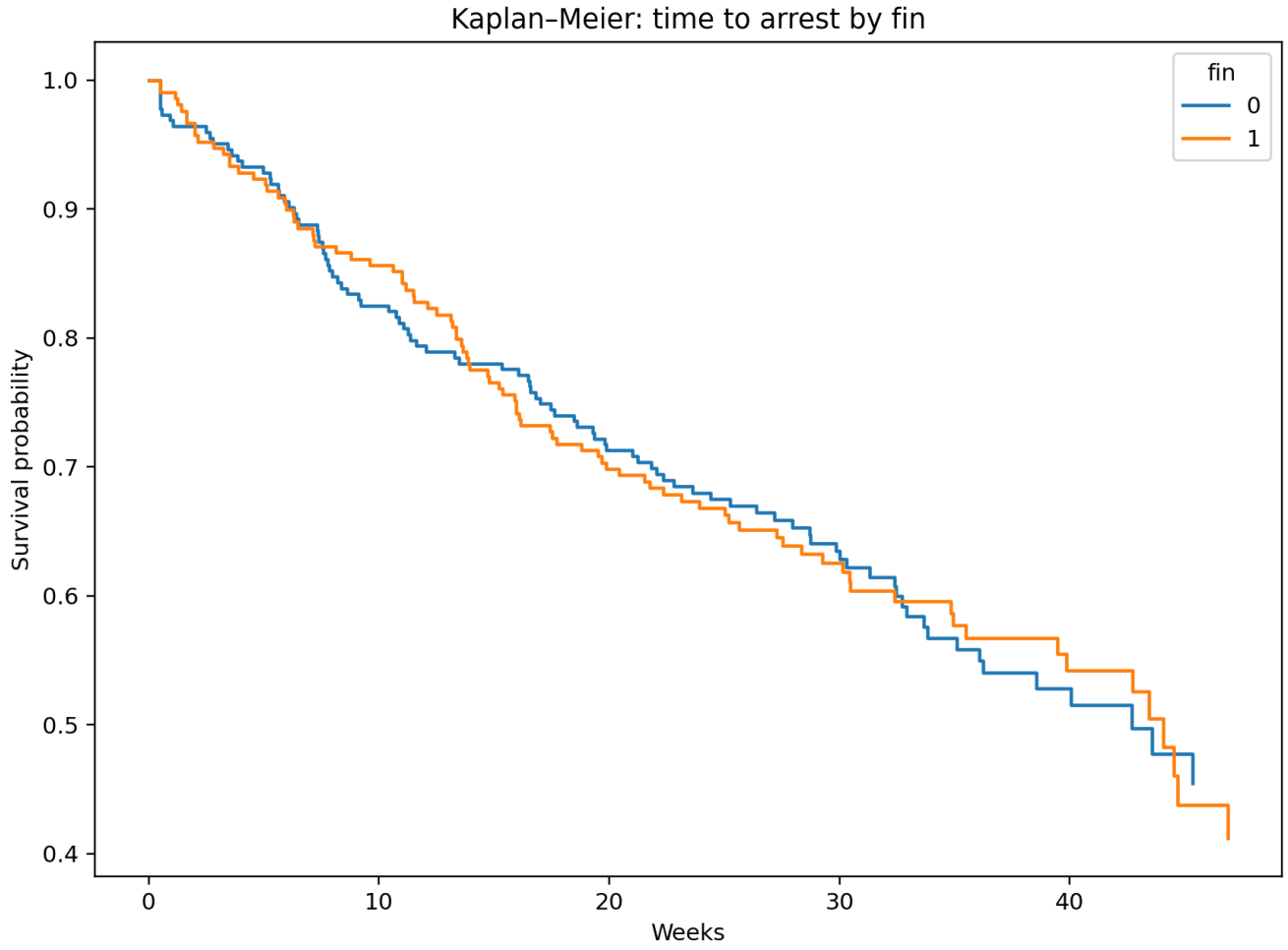
Kaplan–Meier Curves (KM) + Log-Rank

```
p_lr <- function(f, data) {
  s <- survdiff(f, data=data)
  pchisq(s$chisq, df=length(s$n)-1, lower.tail=FALSE)
}

km <- function(var, file) {
  f <- as.formula(paste0("Surv(week, arrest) ~ ", var))
  fit <- survfit(f, data=d)
  pv <- p_lr(f, d)

  g <- ggsurvplot(
    fit, data=d,
    risk.table=TRUE, conf.int=FALSE,
    pval=paste0("p = ", signif(pv, 3)),
    ggtheme=th()
  )

  ggsave(file.path(out, paste0(file, ".png")), g$plot,
         width=8, height=5, dpi=220)

  invisible(g)
}
```

Interpretation: KM curves provide non-parametric survival estimates; the log-rank p-value summarizes whether group differences are statistically detectable over time.

Kaplan–Meier by Financial Aid (fin)

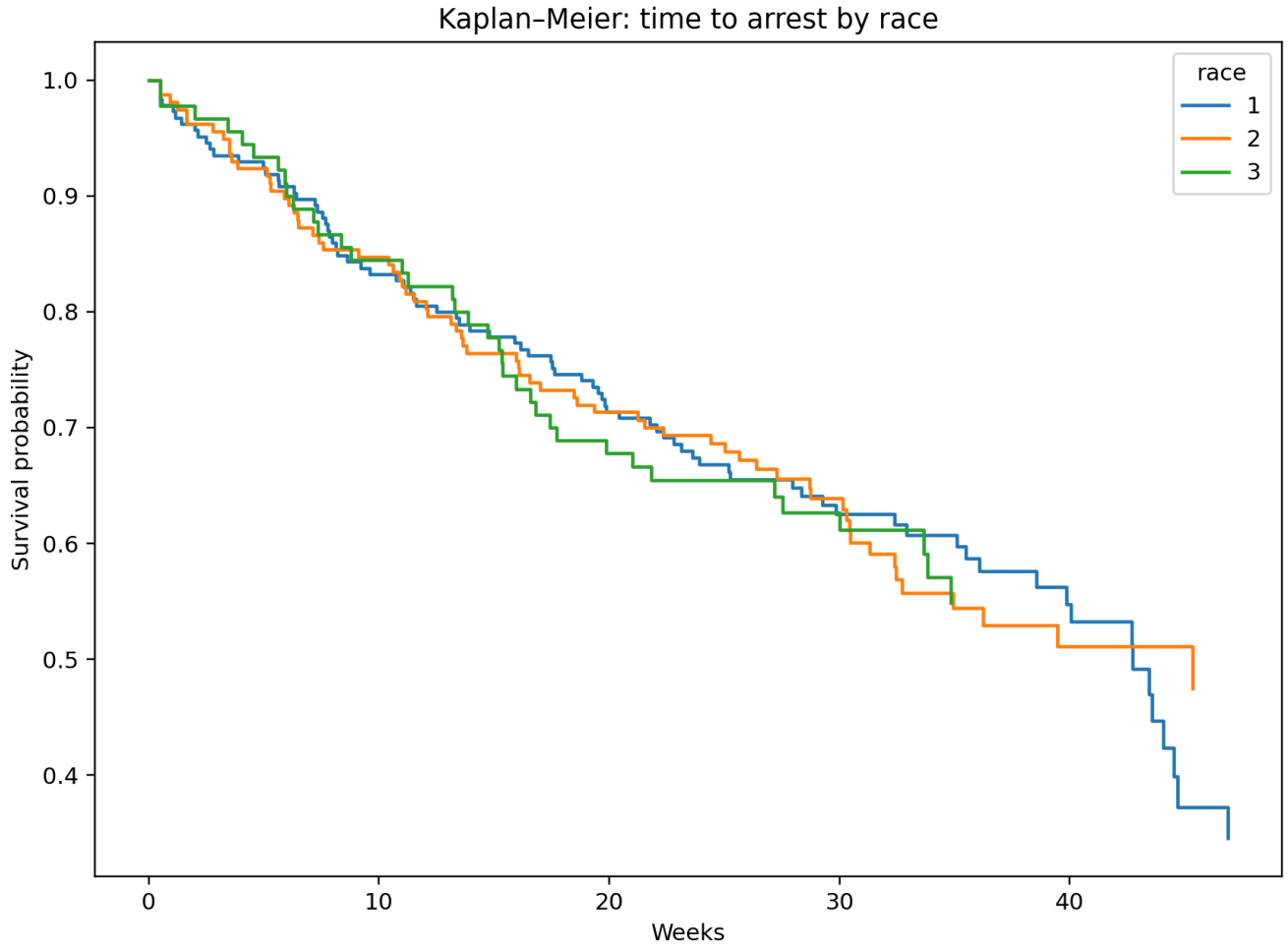
```
km("fin",  "08_km_fin")
```



Interpretation: Visual separation between groups indicates differential event-free probability and motivates covariate inclusion in the Cox model.

Kaplan–Meier by Race (race)

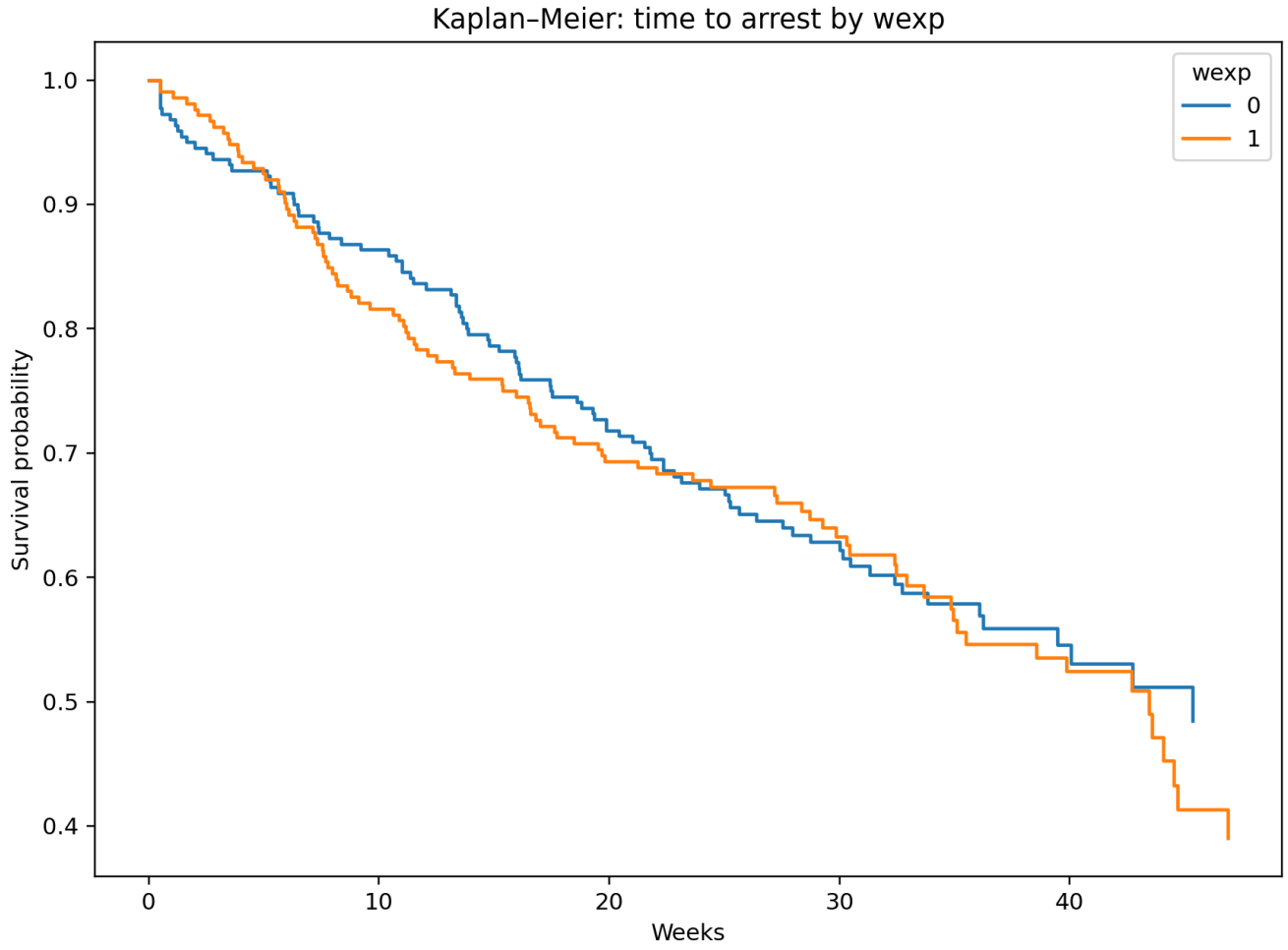
```
km("race", "09_km_race")
```



Interpretation: Visual separation between groups indicates differential event-free probability and motivates covariate inclusion in the Cox model.

Kaplan–Meier by Work Experience (wexp)

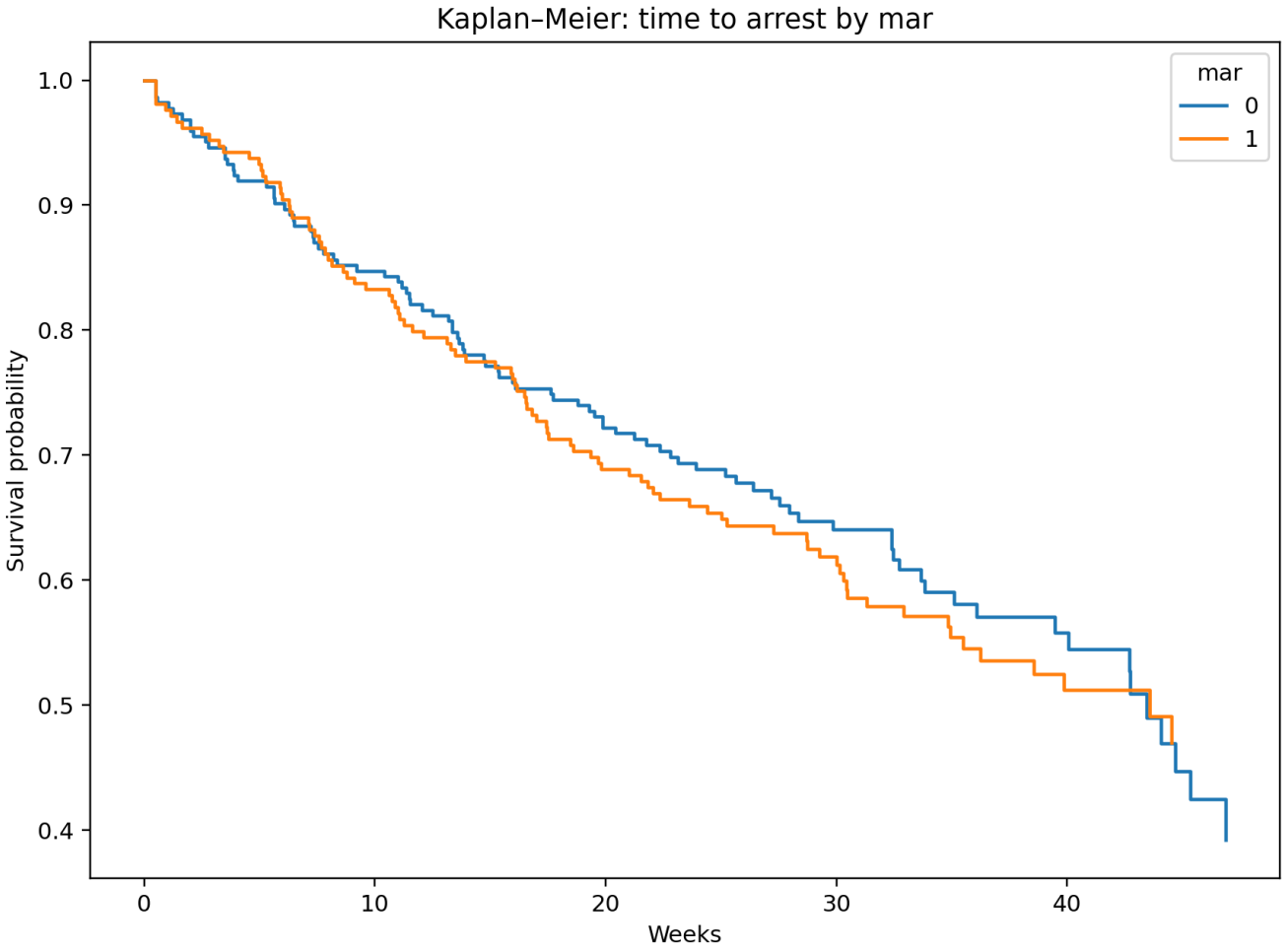
```
km("wexp", "10_km_wexp")
```



Interpretation: Visual separation between groups indicates differential event-free probability and motivates covariate inclusion in the Cox model.

Kaplan–Meier by Marital Status (mar)

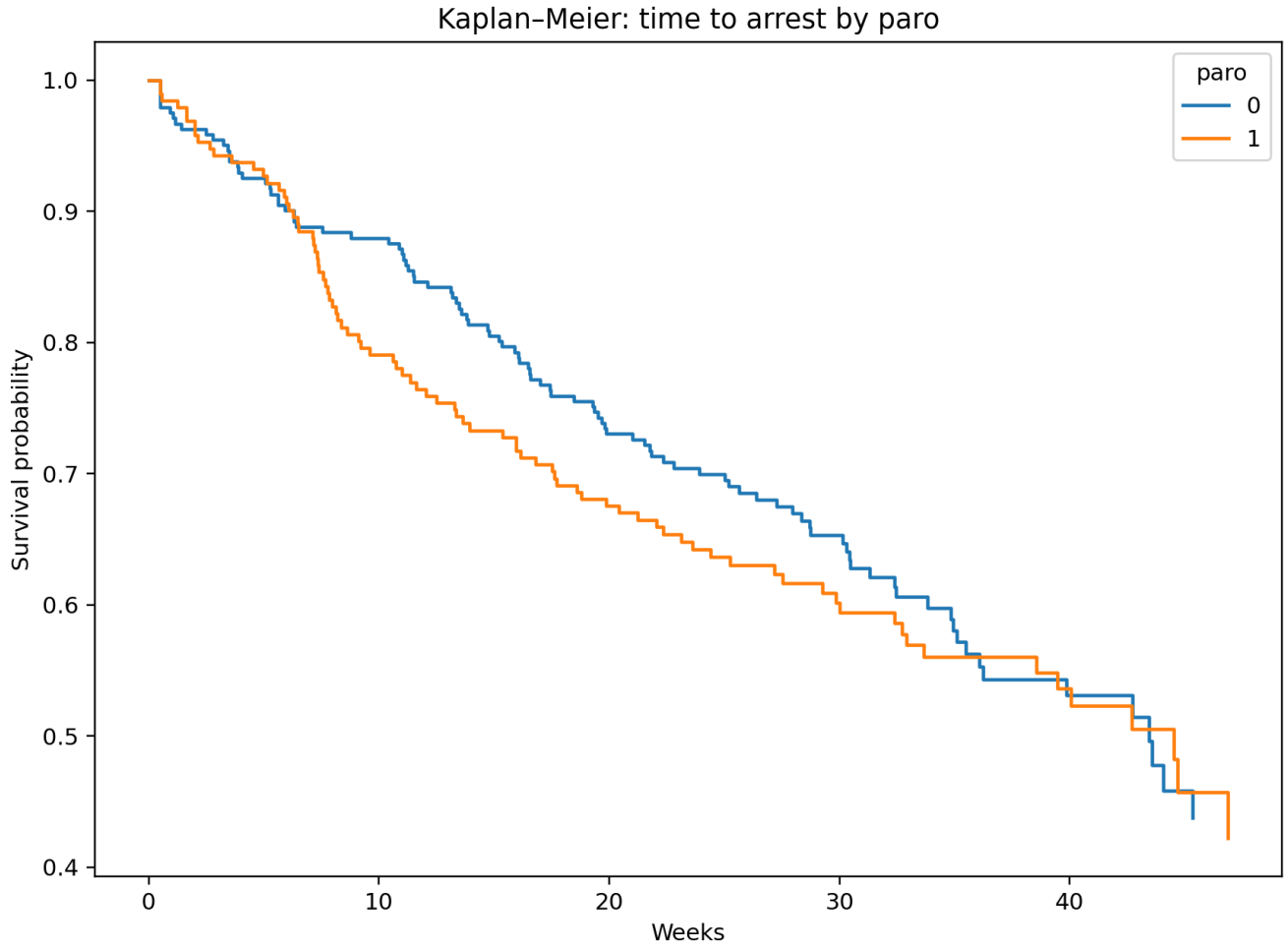
```
km("mar",  "11_km_mar")
```



Interpretation: Visual separation between groups indicates differential event-free probability and motivates covariate inclusion in the Cox model.

Kaplan–Meier by Parole Status (paro)

```
km("paro", "12_km_paro")
```



Interpretation: Visual separation between groups indicates differential event-free probability and motivates covariate inclusion in the Cox model.

Cox Proportional Hazards Model + PH Diagnostics

```
m <- coxph(Surv(week, arrest) ~ fin + wexp + mar + paro + race + educ + prio, data=d)
print(summary(m))

hr <- broom::tidy(m, exponentiate=TRUE, conf.int=TRUE) %>%
  select(term, estimate, conf.low, conf.high, p.value) %>%
  rename(HR=estimate, low=conf.low, high=conf.high, p=p.value)

write.csv(hr, file.path(out, "cox_hr.csv"), row.names=FALSE)

z <- cox.zph(m)
print(z)
capture.output(z, file=file.path(out, "cox_ph.txt"))

png(file.path(out, "cox_ph.png"), width=1200, height=900, res=150)
plot(z)
dev.off()
```

Interpretation: Hazard ratios quantify the direction and magnitude of association with time-to-event, while PH diagnostics validate the proportional hazards assumption.