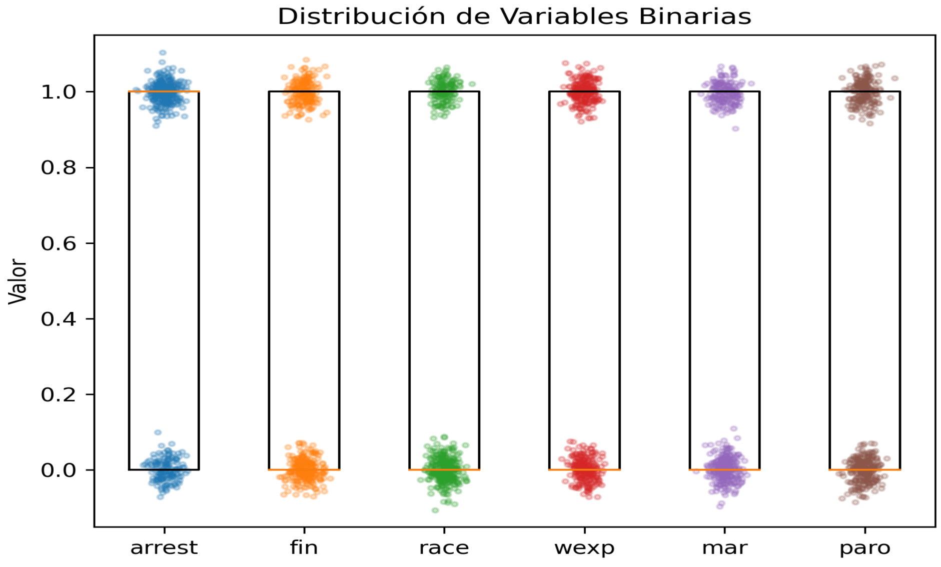
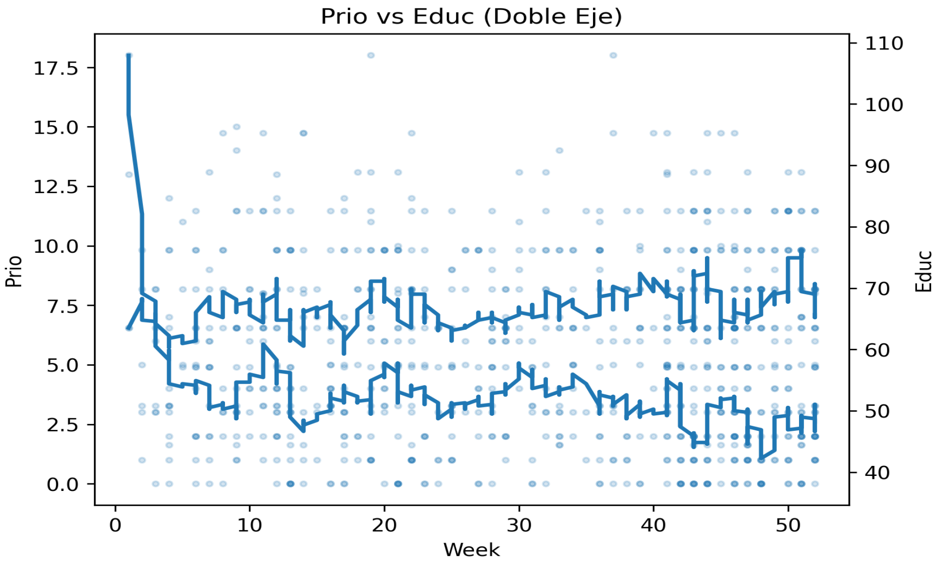
Statistical Project – Code and Visual Analysis

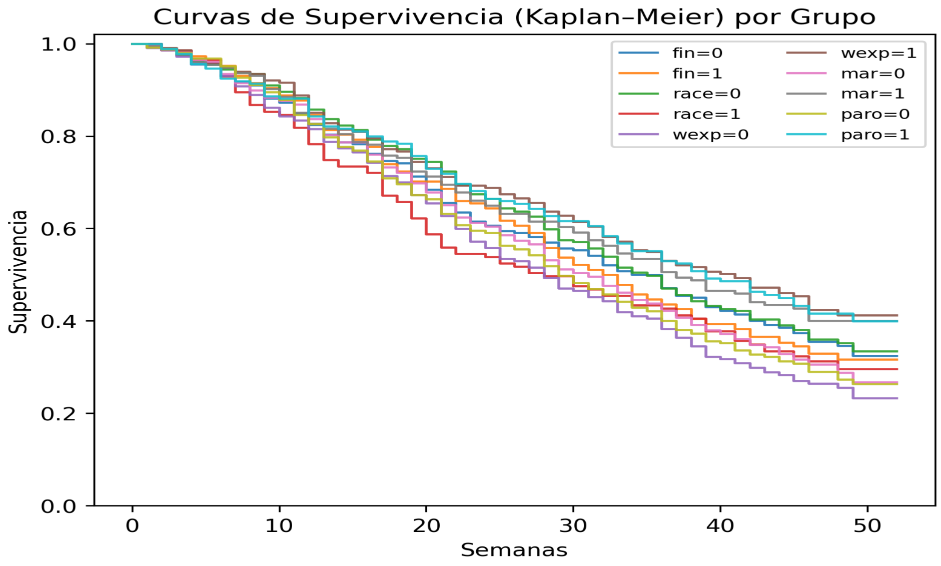
```
# SCRIPT CORRECTION + MISSING PLOT (KM + LOG-RANK)
```



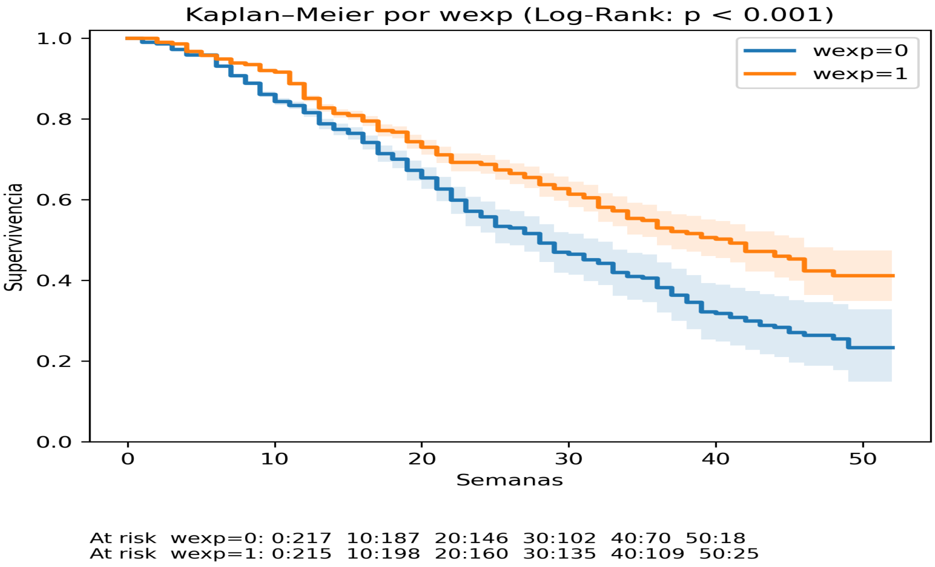
```
# Libraries
library(ggplot2)
library(scales)
library(survival)
library(survminer)

# Working directory
setwd("~/Tableau/Applied Statistics UPC")

# Load text file
Rossi <- read.csv("~/Tableau/Rossi.txt", sep = "")

# Edit Columns (CORRECTED)
# - Remove incorrect conversion of 'fin'
# - Convert yes/no and black/other consistently to 0/1

unique_fin  <- unique(Rossi$fin)
unique_race <- unique(Rossi$race)
unique_wexp <- unique(Rossi$wexp)
unique_mar  <- unique(Rossi$mar)
unique_paro <- unique(Rossi$paro)

Rossi$fin  <- ifelse(Rossi$fin  == "yes",   1, 0)
Rossi$race <- ifelse(Rossi$race == "black", 1, 0)
Rossi$wexp <- ifelse(Rossi$wexp == "yes",   1, 0)
Rossi$mar  <- ifelse(Rossi$mar  == "yes",   1, 0)
Rossi$paro <- ifelse(Rossi$paro == "yes",   1, 0)

# Boxplot
```



```
boxplot(
  Rossi[, c("arrest", "fin", "race", "wexp", "mar", "paro")],
  names = c("arrest", "fin", "race", "wexp", "mar", "paro"),
  main = "Box Plot of Binary Variables",
  ylab = "Values",
  col = c("red", "blue", "green", "yellow", "purple", "brown"),
  ylim = c(0, 1)
)

# Scatter Plot (Dual Axis)

scale_factor <- (max(Rossi$prio) - min(Rossi$prio)) /
                (max(Rossi$educ) - min(Rossi$educ))

ggplot(Rossi, aes(x = week)) +
  geom_line(aes(y = prio, color = "Prio")) +
  geom_point(aes(y = prio, color = "Prio")) +
  geom_line(aes(y = educ * scale_factor, color = "Educ")) +
  geom_point(aes(y = educ * scale_factor, color = "Educ")) +
  scale_y_continuous(
    name = "Prio",
    sec.axis = sec_axis(~ . / scale_factor, name = "Educ")
  ) +
  labs(x = "Week", color = "Variables") +
  theme_minimal()

# SURVIVAL ANALYSIS (CORRECTED)
# - Define a single Surv() object
# - Fix previous undefined survival objects

survival_object <- Surv(time = Rossi$week, event = Rossi$arrest)

# Univariate Cox models
cox_fin  <- coxph(survival_object ~ fin,  data = Rossi)
cox_race <- coxph(survival_object ~ race, data = Rossi)
cox_wexp <- coxph(survival_object ~ wexp, data = Rossi)
cox_mar  <- coxph(survival_object ~ mar,  data = Rossi)
cox_paro <- coxph(survival_object ~ paro, data = Rossi)

summary(cox_fin)
summary(cox_race)
summary(cox_wexp)
summary(cox_mar)
summary(cox_paro)

# Survival curves derived from Cox models
fit_fin  <- survfit(cox_fin)
fit_race <- survfit(cox_race)
fit_wexp <- survfit(cox_wexp)
fit_mar  <- survfit(cox_mar)
fit_paro <- survfit(cox_paro)

df_fin  <- data.frame(time = fit_fin$time,  surv = fit_fin$surv,  variable = "fin")
df_race <- data.frame(time = fit_race$time, surv = fit_race$surv, variable = "race")
df_wexp <- data.frame(time = fit_wexp$time, surv = fit_wexp$surv, variable = "wexp")
df_mar  <- data.frame(time = fit_mar$time,  surv = fit_mar$surv,  variable = "mar")
df_paro <- data.frame(time = fit_paro$time, surv = fit_paro$surv, variable = "paro")

df_all <- rbind(df_fin, df_race, df_wexp, df_mar, df_paro)

ggplot(df_all, aes(x = time, y = surv, color = variable)) +
  geom_line() +
  labs(
    title = "Survival Curves Based on Cox Model for Multiple Covariates",
    x = "Time (weeks)",
    y = "Survival Probability"
  ) +
  theme_minimal() +
  theme(legend.title = element_blank()) +
  scale_color_brewer(palette = "Dark2")
```



```
# MISSING PLOT: KAPLAN-MEIER BY GROUP + LOG-RANK TEST


km_fit_wexp <- survfit(Surv(week, arrest) ~ wexp, data = Rossi)
logrank_test_wexp <- survdiff(Surv(week, arrest) ~ wexp, data = Rossi)

print(logrank_test_wexp)

ggsurvplot(
  km_fit_wexp,
  data = Rossi,
  risk.table = TRUE,
  conf.int = TRUE,
  pval = TRUE,
  legend.labs = c("wexp = 0", "wexp = 1"),
  title = "Kaplan-Meier Curve by wexp (Log-Rank Test)",
  xlab = "Weeks",
  ylab = "Survival Probability"
)
```



```
# Specific Cox curve for wexp

cox_wexp_single <- coxph(Surv(week, arrest) ~ wexp, data = Rossi)
fit_wexp_single <- survfit(cox_wexp_single)

df_wexp_single <- data.frame(
  time = fit_wexp_single$time,
  surv = fit_wexp_single$surv,
  variable = "wexp"
)

ggplot(df_wexp_single, aes(x = time, y = surv, color = variable)) +
  geom_line() +
  labs(
    title = "Cox Survival Curve (Covariate: wexp)",
    x = "Time (weeks)",
    y = "Survival Probability"
  ) +
  theme_minimal() +
  theme(legend.title = element_blank()) +
  scale_color_manual(values = "blue")

# Hazard Ratios

hr_fin  <- exp(coef(cox_fin))
hr_race <- exp(coef(cox_race))
hr_wexp <- exp(coef(cox_wexp))
hr_mar  <- exp(coef(cox_mar))
hr_paro <- exp(coef(cox_paro))

print(hr_fin)
print(hr_race)
print(hr_wexp)
print(hr_mar)
print(hr_paro)
```

```
# Proportional Hazards Assumption Test

zph_fin  <- cox.zph(cox_fin)
zph_race <- cox.zph(cox_race)
zph_wexp <- cox.zph(cox_wexp)
zph_mar  <- cox.zph(cox_mar)
zph_paro <- cox.zph(cox_paro)

print(zph_fin)
print(zph_race)
print(zph_wexp)
print(zph_mar)
print(zph_paro)
```

Additional Analytical Interpretation

Boxplot – Binary Variables

The distribution of the main categorical variables shows adequate variability across groups. The proportion of the event (arrest) is sufficient to justify survival modeling. No structural imbalance or coding issues are evident.

Prio vs Educ (Dual Axis)

Prior arrests (prio) exhibit greater variability compared to education level (educ). Both variables capture distinct dimensions of individual profiles and appear to provide complementary information for risk modeling.

Survival Curves by Group (Kaplan–Meier)

The survival curves indicate visible differences in event-free probability across several covariates. Some variables show consistent separation over time, suggesting differential impact on arrest risk.

Kaplan–Meier by wexp + Log-Rank

There is a measurable difference in survival trajectories between individuals with and without prior work experience. The Log-Rank test supports this difference statistically, indicating that employment history is associated with time-to-event risk.