Statistical Packages

HW8

Mohamed Salem

```
# download the files
library(downloader)
download("http://databank.worldbank.org/data/download/Edstats_csv.zip",
    dest = "Edstats_csv.zip")
unzip("Edstats_csv.zip", exdir = ".")
# read the data
EdStatsData <- read.csv("./Edstats_csv/EdStatsData.csv", stringsAsFactors = FALSE)
# Rename the data
nameph <- gsub("X", "", names(EdStatsData), fixed = TRUE)
nameph[1:4] <- c("Country_Name", "Country_Code", "Indicator_Name",
    "Indicator_Code")
names(EdStatsData) <- nameph
EdStatsData[, 70] <- NULL
EdStatsData[, c(2, 4)] <- NULL
EdStatsData_cln <- gather(EdStatsData, Year, Value, 3:67)
```

The data previously consisted of $\sim$ 900k obsrevations of 70 variables. Now, it consists of $\sim$ 58 million observations of 4 variables.

```
jpvus <- filter(EdStatsData_cln, (Country_Name == "Japan" | Country_Name ==
    "United States") & Indicator_Name == "GDP per capita (current US$)")
jpvus_wide <- spread(jpvus, Country_Name, Value)
# Below code adapted from
# https://www.r-bloggers.com/example-8-41-scatterplot-with-marginal-histograms/
scatterhist = function(x, y, xlab = "", ylab = "") {
    zones = matrix(c(2, 0, 1, 3), ncol = 2, byrow = TRUE)
    layout(zones, widths = c(4/5, 1/5), heights = c(1/5, 4/5))
    xhist = hist(x, plot = FALSE)
    yhist = hist(y, plot = FALSE)
    top = max(c(xhist$counts, yhist$counts))
    par(mar = c(3, 3, 1, 1))
    plot(x, y)
    abline(a = 0, b = 1)
    par(mar = c(0, 3, 1, 1))
    barplot(xhist$counts, axes = FALSE, ylim = c(0, top), space = 0)
    par(mar = c(3, 0, 1, 1))
    barplot(yhist$counts, axes = FALSE, xlim = c(0, top), space = 0,
        horiz = TRUE)
    par(oma = c(3, 3, 0, 0))
    mtext(xlab, side = 1, line = 2, outer = TRUE, adj = 0.35, at = 1.5 *
        (mean(x) - min(x))/(max(x) - min(x)))
    mtext(ylab, side = 2, line = 2, outer = TRUE, adj = 0.35, at = (0.1 *
        (mean(y) - min(y))/(max(y) - min(y))))
}
scatterhist(jpvus_wide$`United States`, jpvus_wide$Japan, xlab = "U.S",
    ylab = "Japan")
```

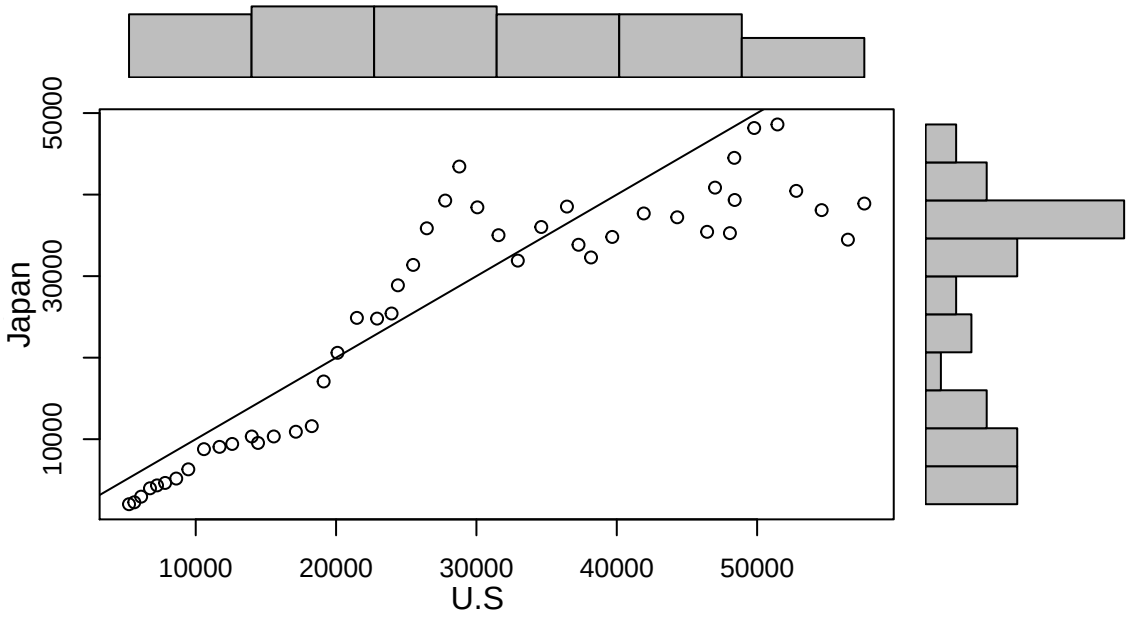

Figure 1: GDP per Capita

```
# Below code adapted from
# https://stackoverflow.com/questions/8545035/scatterplot-with-marginal-histograms-in-ggplot2
hist_top <- ggplot() + geom_histogram(aes(jpvus_wide$`United States`),
    binwidth = 5000, color = "white") + labs(x = "", y = "") + theme(plot.margin = margin(0.2,
    0.2, 0.2, 0.2, "cm"), axis.ticks.y = element_blank(), axis.text.y = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_blank())
empty <- ggplot() + geom_point(aes(0, 0), colour = "white") + theme(axis.ticks = element_blank(),
    panel.background = element_blank(), axis.text.x = element_blank(),
    axis.text.y = element_blank(), axis.title.x = element_blank(),
    axis.title.y = element_blank())

scatter <- ggplot() + geom_point(aes(jpvus_wide$`United States`, jpvus_wide$Japan)) +
    labs(x = "U.S", y = "Japan") + geom_abline(intercept = 0, slope = 1) +
    theme(plot.margin = margin(0.2, 0.2, 0.2, 0.2, "cm"))
hist_right <- ggplot() + geom_histogram(aes(jpvus_wide$Japan), binwidth = 5000,
    color = "white") + coord_flip() + labs(x = "", y = "") + theme(plot.margin = margin(0.2,
    0.2, 0.2, 0.2, "cm"), axis.ticks.y = element_blank(), axis.text.y = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_blank())
grid.arrange(hist_top, empty, scatter, hist_right, ncol = 2, nrow = 2,
    widths = c(3, 1), heights = c(1, 3))
```

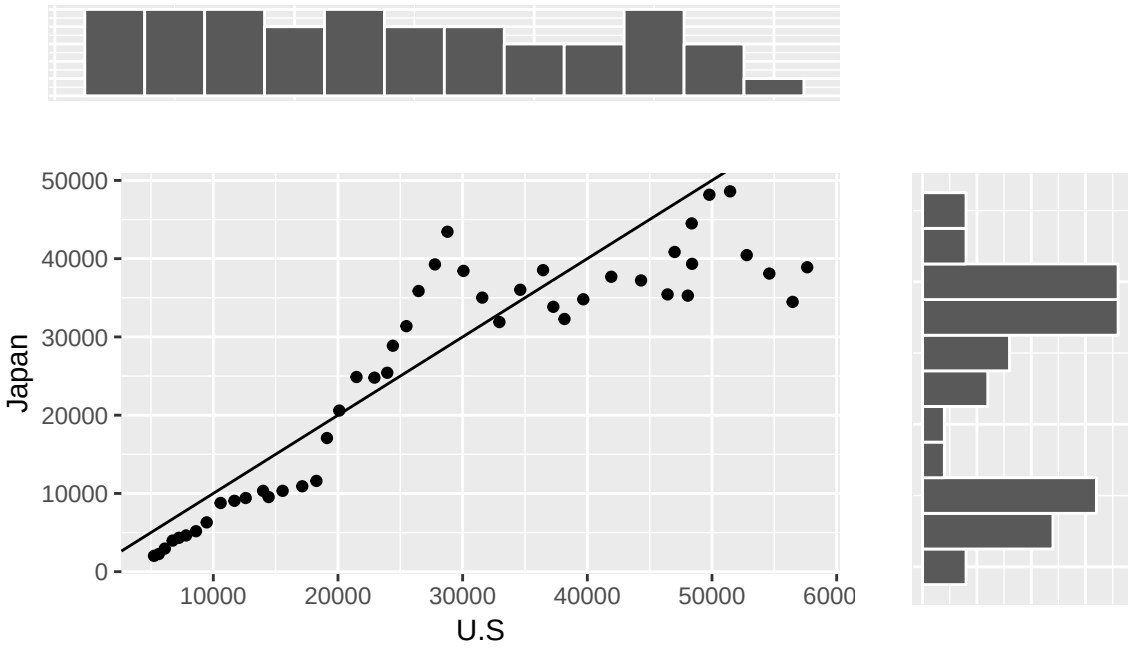

Figure 2: GDP per Capita