Descriptive Statistics — iFood CSV (DOCX-safe)

2025-10-15

Contents

1 Reference Descriptive Script (DOCX-safe)

Note: Use forward slashes in paths on Windows (e.g., `C:/Users/...`).
Place this `.Rmd` in the same folder as `ifood_base.csv` or update `data_path` below.

This version avoids HTML-only features so it **knits to Word (.docx) without errors**. If you knit to HTML, it will still show a floating ToC and nicer tables.

```
required_pkgs <- c("readr","dplyr","tidyr","stringr","purrr","lubridate","knitr")
to_install <- setdiff(required_pkgs, rownames(installed.packages()))
if (length(to_install)) install.packages(to_install, repos = "https://cloud.r-project.org")

library(readr); library(dplyr); library(tidyr); library(stringr); library(purrr)
library(lubridate); library(knitr)

# Safe Windows path
safe_path <- function(p) ifelse(is.na(p) || !nzchar(p), p, gsub("\\\\", "/", p))

# Table helper: uses HTML styling only when knitting to HTML; plain kable for Word/PDF
theme_table <- function(tbl) {
  if (knitr::is_html_output()) {
```

```
    if (!requireNamespace("kableExtra", quietly = TRUE)) {
      return(knitr::kable(tbl, align = "l"))
    }
    kableExtra::kbl(tbl, booktabs = TRUE, align = "l") |>
      kableExtra::kable_styling(full_width = FALSE, bootstrap_options = c("striped","hover","c
  } else {
    knitr::kable(tbl, align = "l")
  }
}

# Plot colors (enough for most categorical vars)
listOfColors <- grDevices::rainbow(40)
```

1.1 Read data

```
# If your CSV is elsewhere, replace with an absolute path using forward slashes
# e.g., data_path <- "ifood_base.csv"
data_path <- "ifood_enriched.csv"
data_path <- safe_path(data_path)

if (!file.exists(data_path)) {
  stop(paste0("CSV not found at: '", data_path, "'. ",
              "Check your username or move the file to this folder."))
}

dd <- readr::read_csv(data_path, show_col_types = FALSE)
```

```
# Keep only selected variables
vars_keep <- c("Education","MaritalSts","Income","Kidhome","Teenhome",
               "Recency","Response","Complain","Age")

dd <- dd %>% select(any_of(vars_keep))

dim(dd)
```

```
## [1] 2031    9
```

```
names(dd)
```

```
## [1] "Education"  "MaritalSts" "Income"     "Kidhome"    "Teenhome"
## [6] "Recency"    "Response"   "Complain"   "Age"
```

1.2 Dimensions and variable names

```
dim(dd)
```

```
## [1] 2031    9
```

```
n <- nrow(dd); K <- ncol(dd)
n; K
```

```
## [1] 2031
```

```
## [1] 9
```

```
names(dd)
```

```
## [1] "Education"  "MaritalSts" "Income"     "Kidhome"    "Teenhome"
## [6] "Recency"    "Response"   "Complain"   "Age"
```

1.3 Optional: Declare/standardize types (recommended for iFood)

```
# Normalize common alternative names
dd <- dd %>% rename(
  Marital_Status= dplyr::any_of(c("Marital_Status","Marital","marital_status")),
  Education     = dplyr::any_of(c("Education","education"))
)

# Parse date column if present
if ("Dt_Customer" %in% names(dd)) {
  dd$Dt_Customer <- suppressWarnings(parse_date_time(dd$Dt_Customer,
                                                     orders = c("dmy","ymd","mdy","d-b-Y","Y-m-
  dd$Dt_Customer <- as.Date(dd$Dt_Customer)
}

# Categorical text columns
for (col in c("Education","Marital_Status")) {
  if (col %in% names(dd)) dd[[col]] <- as.factor(dd[[col]])
}

# Binary 0/1 columns (if present)
bin_cols <- intersect(c("AcceptedCmp1","AcceptedCmp2","AcceptedCmp3","AcceptedCmp4",
                        "AcceptedCmp5","Complain","Response"), names(dd))
for (bc in bin_cols) {
  if (all(na.omit(unique(dd[[bc]])) %in% c(0,1))) {
```

```
    dd[[bc]] <- factor(dd[[bc]], levels = c(0,1))
  }
}

str(dd)
```

```
## tibble [2,031 x 9] (S3: tbl_df/tbl/data.frame)
##  $ Education : Factor w/ 5 levels "2n Cycle","Basic",..: 3 5 5 5 5 1 3 4 3 4 ...
##  $ MaritalSts: chr [1:2031] "Single" "Together" "Single" "Divorced" ...
##  $ Income    : num [1:2031] 58138 30351 82800 46610 48948 ...
##  $ Kidhome   : num [1:2031] 0 1 0 0 0 0 0 0 0 0 ...
##  $ Teenhome  : num [1:2031] 0 0 0 2 0 0 0 0 0 0 ...
##  $ Recency   : num [1:2031] 58 19 23 8 53 24 54 55 30 12 ...
##  $ Response  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Complain  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age       : num [1:2031] 63 46 74 68 77 24 43 65 54 38 ...
```

1.4 Missingness overview

```
miss_tbl <- tibble::tibble(
  variable = names(dd),
  missing_n = sapply(dd, function(x) sum(is.na(x))),
  missing_pct = round(100 * sapply(dd, function(x) mean(is.na(x))), 2)
) %>% arrange(desc(missing_pct))

theme_table(miss_tbl)
```

variable	missing_n	missing_pct
Education	0	0
MaritalSts	0	0
Income	0	0
Kidhome	0	0
Teenhome	0	0
Recency	0	0
Response	0	0
Complain	0	0
Age	0	0

1.5 Descriptive function

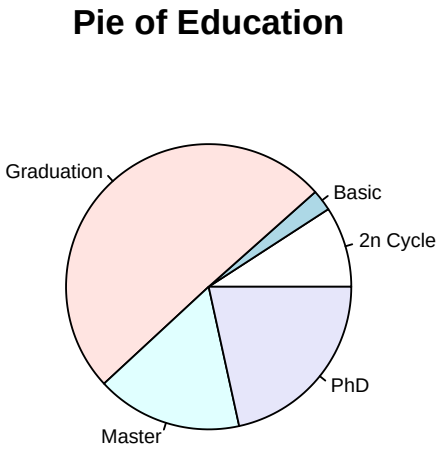
```
descriptiva <- function(X, nom, nrows_total) {
  if (!(is.numeric(X) || inherits(X, "Date"))) {
```

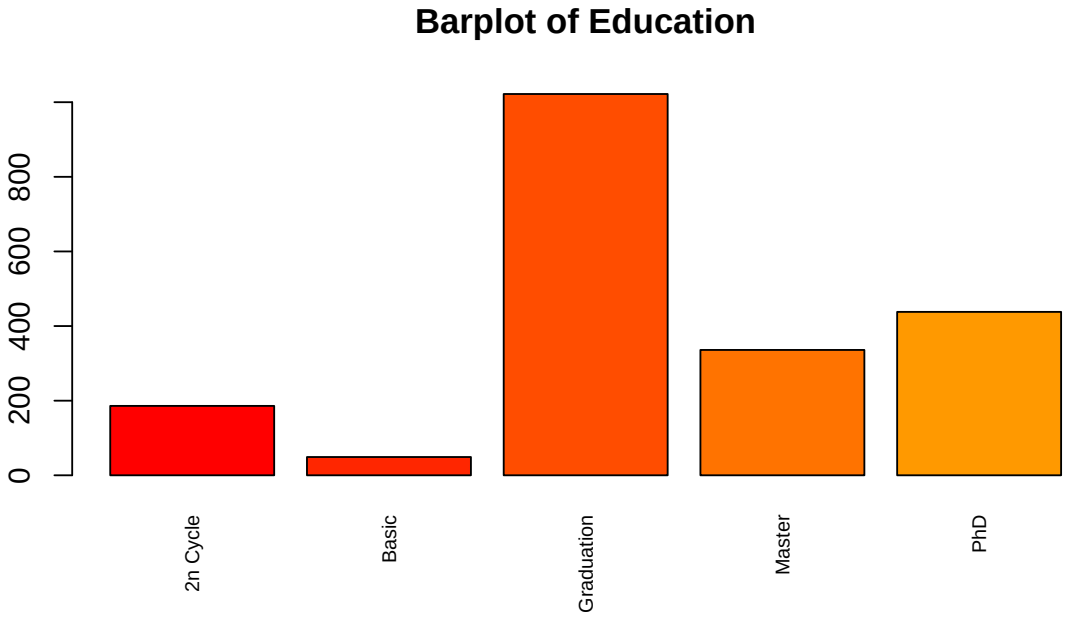
```
    # Categorical
    frecs <- table(as.factor(X), useNA = "ifany")
    proportions <- frecs / nrows_total
    pie(frecs, cex = 0.7, main = paste("Pie of", nom))
    barplot(frecs, las = 3, cex.names = 0.7,
            main = paste("Barplot of", nom), col = listOfColors)
    cat("\nNumber of categories:", length(frecs), "\n")
    cat("\nFrequency table\n"); print(frecs)
    cat("\nRelative frequency table\n"); print(round(proportions, 4))
  } else {
    # Numeric
    hist(X, main = paste("Histogram of", nom), xlab = nom)
    boxplot(X, horizontal = TRUE, main = paste("Boxplot of", nom), xlab = nom)
    cat("\nSummary Statistics for", nom, "\n"); print(summary(X))
    sd_x <- sd(X, na.rm = TRUE)
    mn_x <- mean(X, na.rm = TRUE)
    cat("\nStandard deviation:", round(sd_x, 4), "\n")
    cat("Coefficient of variation (sd/mean):",
        ifelse(is.na(mn_x) || mn_x == 0, NA, round(sd_x / mn_x, 4)), "\n")
  }
}
```

1.6 Run basic descriptives for all variables

```
for (col in names(dd)) {
  cat("\n\n## Variable:", col, "\n")
  descriptiva(dd[[col]], col, nrow(dd))
}
```

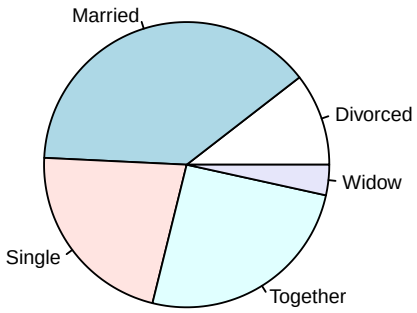
```
##
##
## ## Variable: Education
```

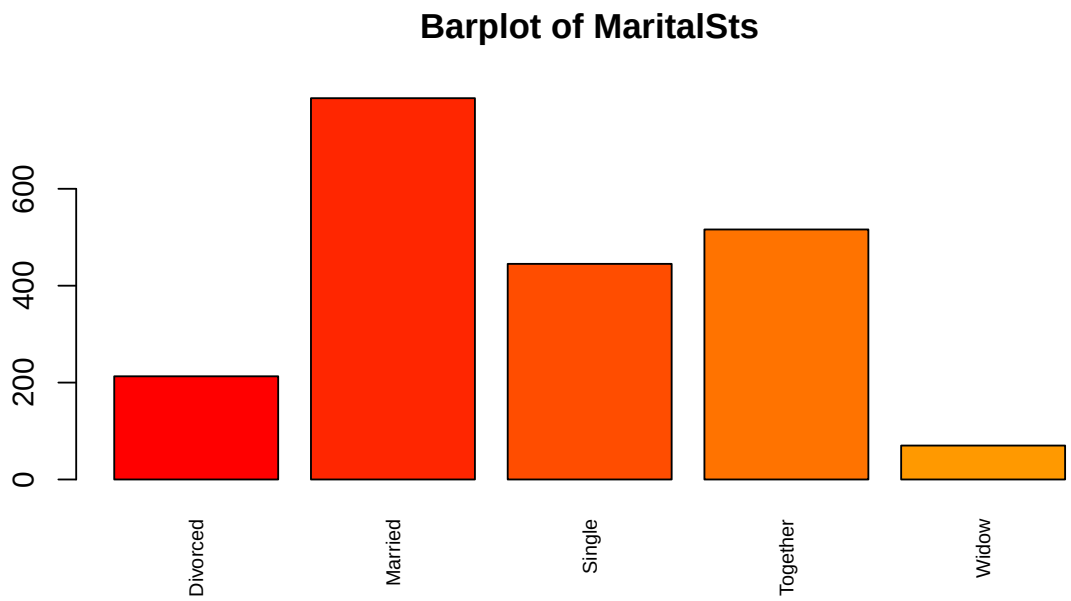




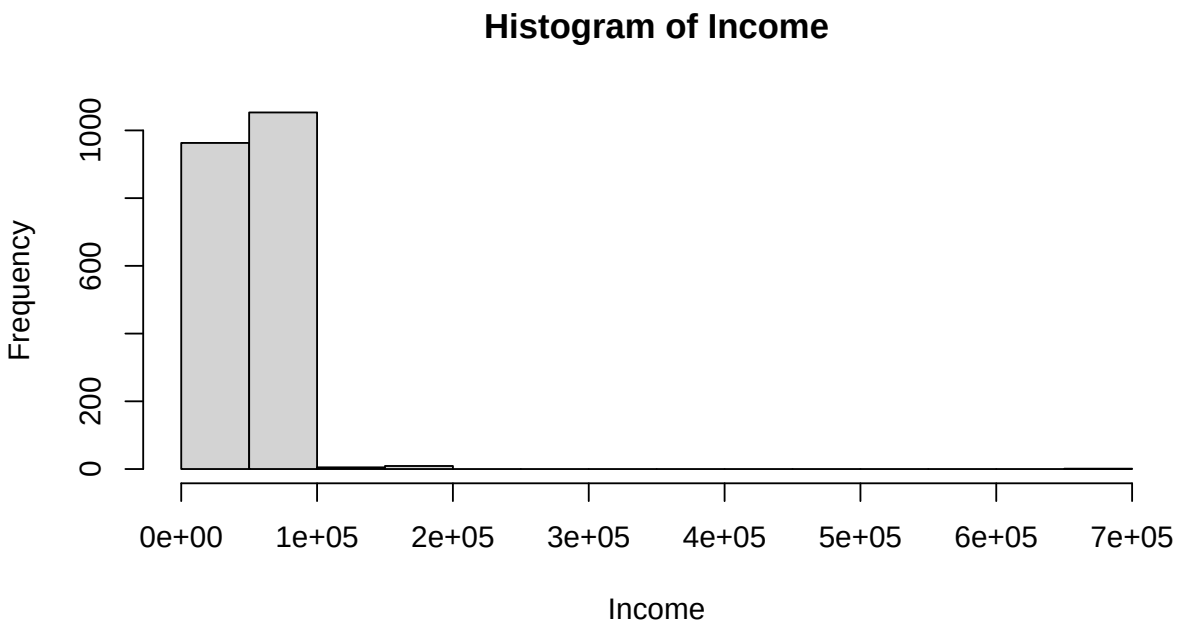

```
##
## Number of categories: 5
##
## Frequency table
##
##   2n Cycle      Basic Graduation     Master        PhD
##        186         49       1022        336        438
```

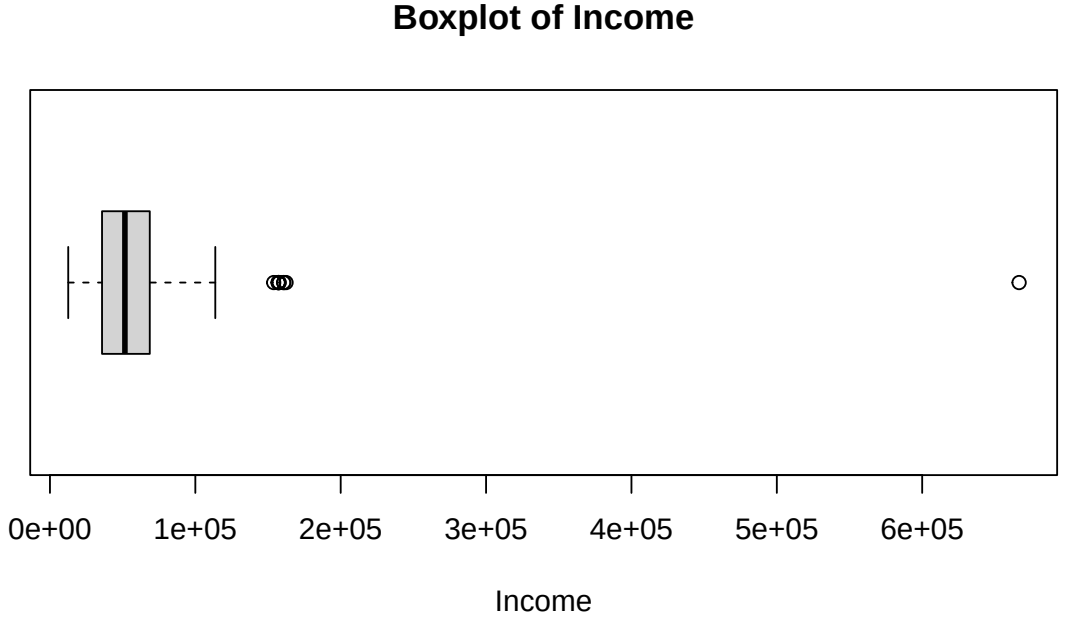
```
##
## Relative frequency table
##
##   2n Cycle      Basic Graduation     Master        PhD
##     0.0916     0.0241     0.5032     0.1654     0.2157
##
##
## ## Variable: MaritalSts
```





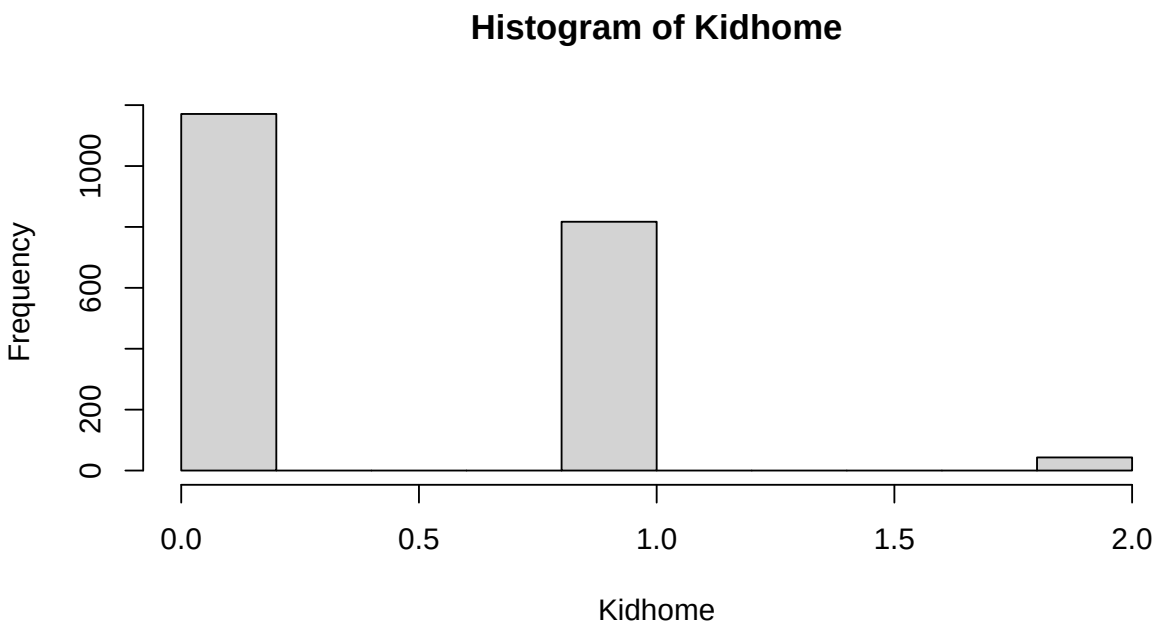

```
##
## Number of categories: 5
##
## Frequency table
##
## Divorced  Married   Single Together    Widow
##      213      787      445      516       70
##
## Relative frequency table
##
## Divorced  Married   Single Together    Widow
##   0.1049   0.3875   0.2191   0.2541   0.0345
##
##
## ## Variable: Income
```

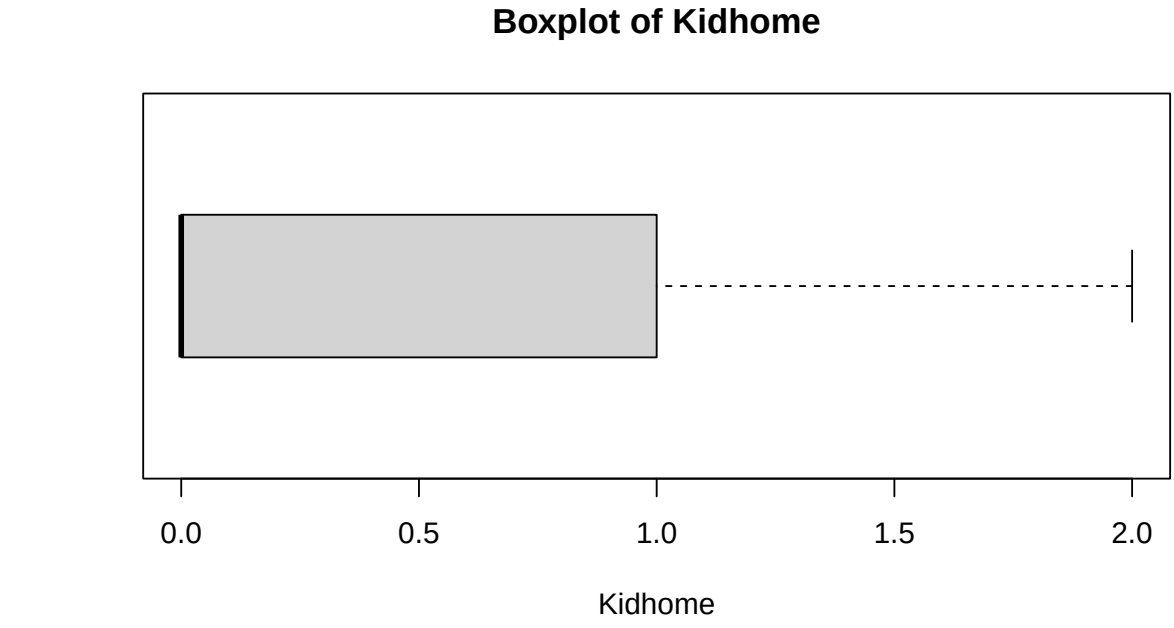





```
##
## Summary Statistics for Income
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12571   35829   51563   52844   68656  666666
##
## Standard deviation: 25242.31
## Coefficient of variation (sd/mean): 0.4777
```

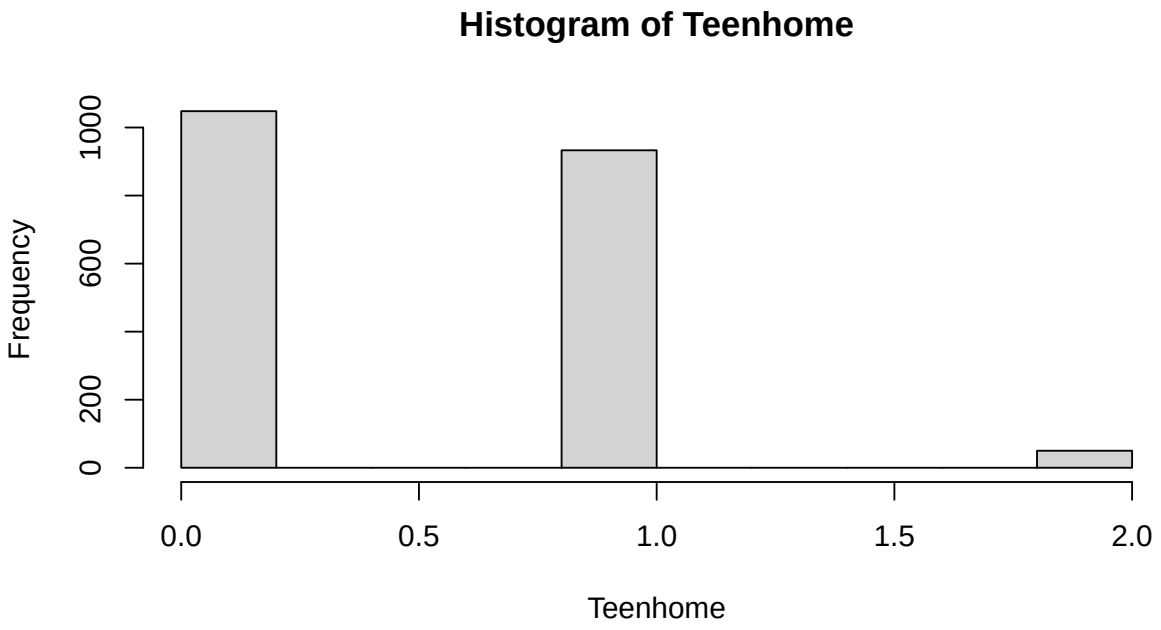
```
##
##
## ## Variable: Kidhome
```

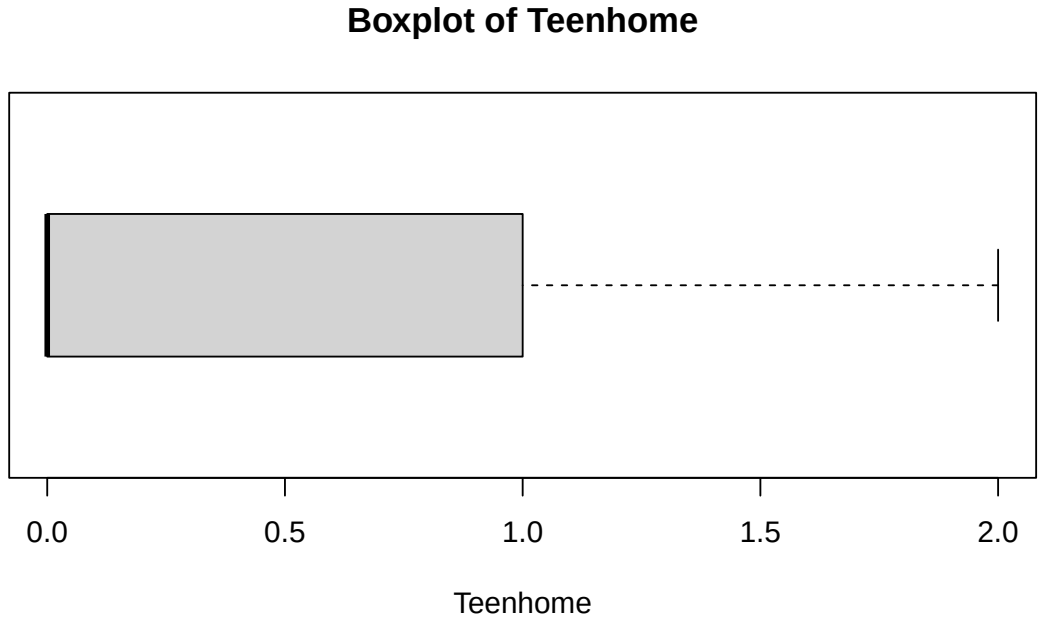




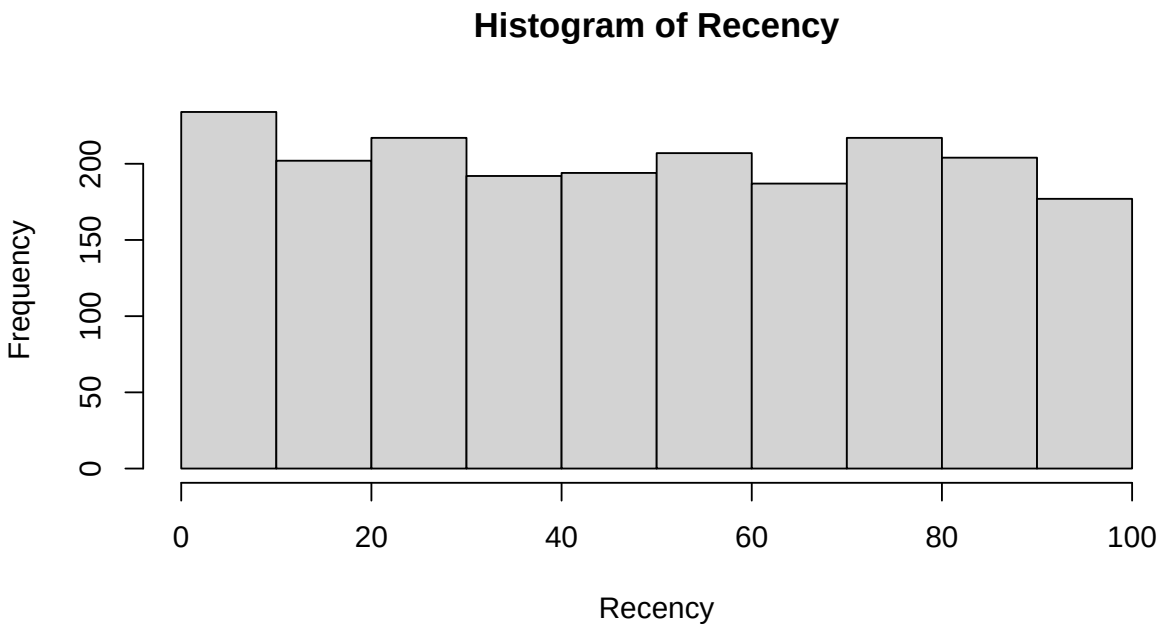

```
##
## Summary Statistics for Kidhome
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

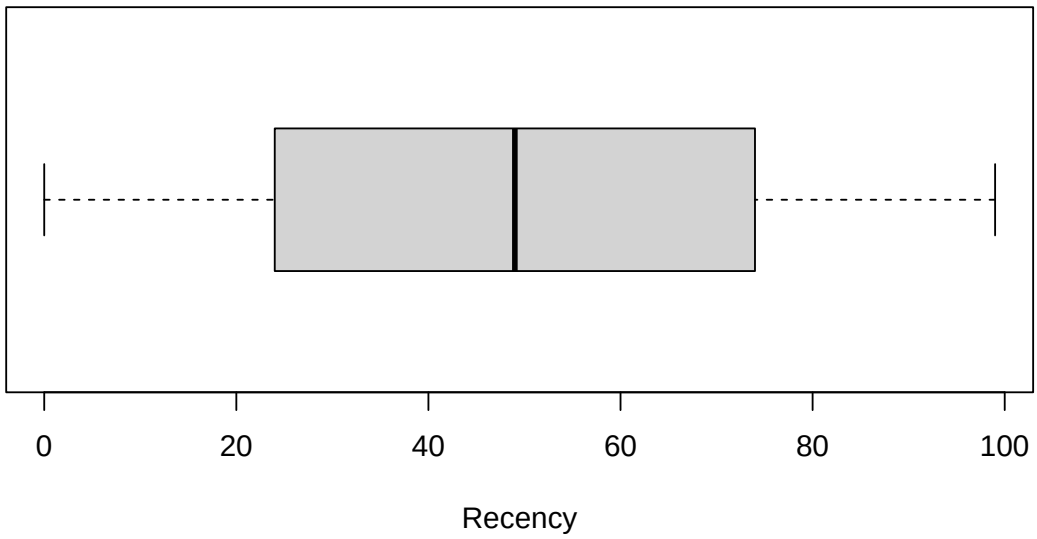
```
##  0.0000  0.0000  0.0000  0.4446  1.0000  2.0000
##
## Standard deviation: 0.538
## Coefficient of variation (sd/mean): 1.21
##
##
## ## Variable: Teenhome
```



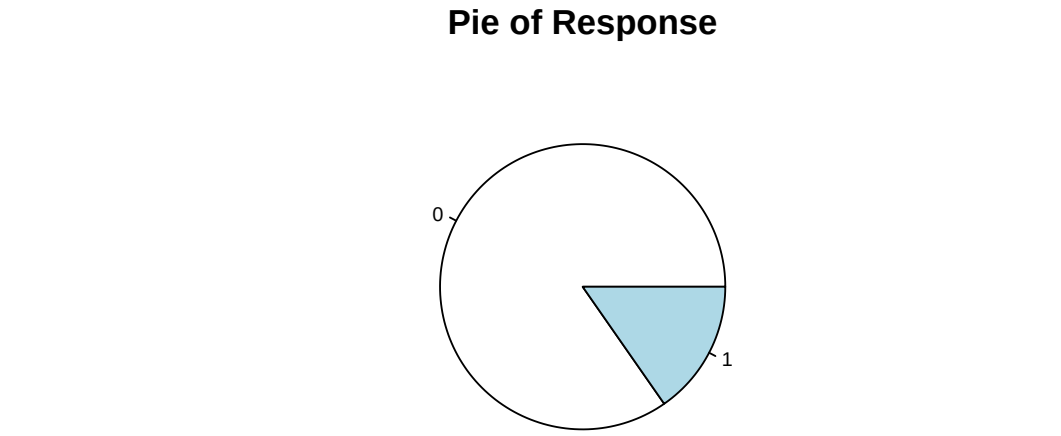


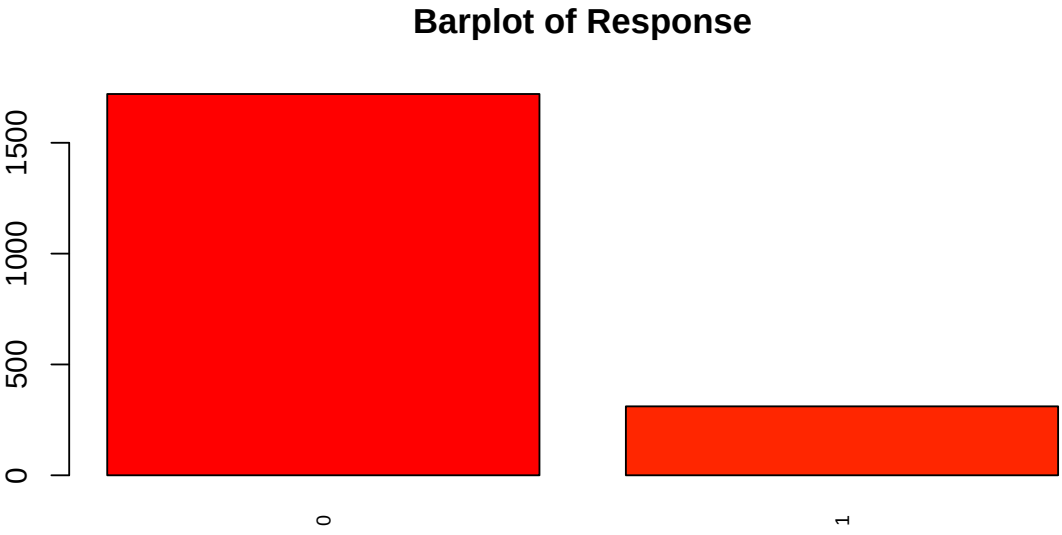
```
##
## Summary Statistics for Teenhome
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.5086  1.0000  2.0000
##
## Standard deviation: 0.5471
## Coefficient of variation (sd/mean): 1.0756
##
##
## ## Variable: Recency
```





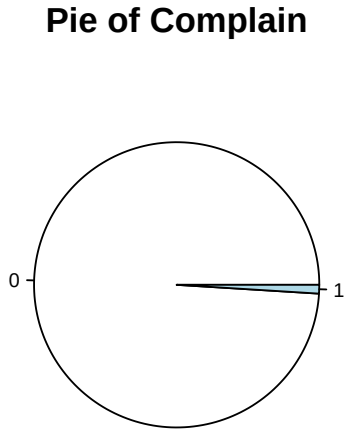

```
##
## Summary Statistics for Recency
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   24.00   49.00   49.14   74.00   99.00
##
## Standard deviation: 28.9768
## Coefficient of variation (sd/mean): 0.5897
##
##
## ## Variable: Response
```

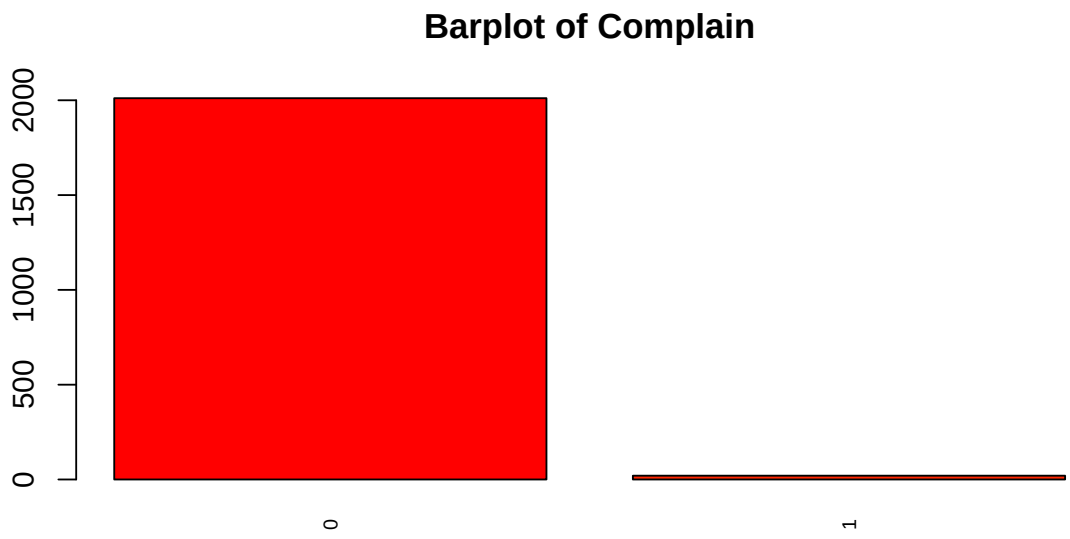




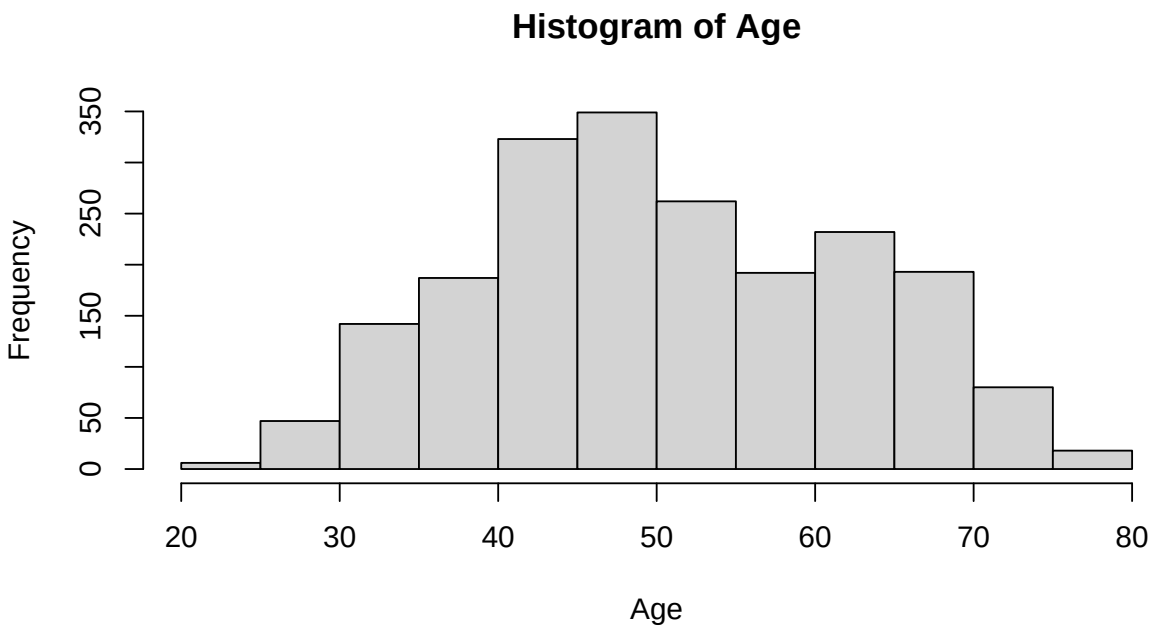

```
##
## Number of categories: 2
##
## Frequency table
##
##    0    1
## 1720  311
```

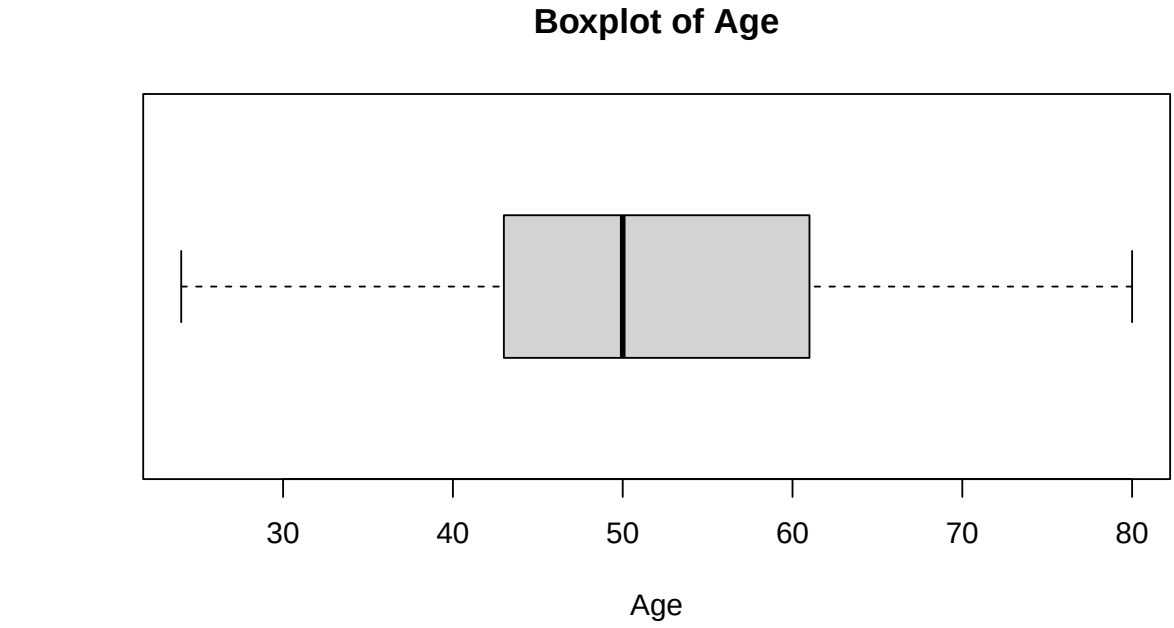
```
##
## Relative frequency table
##
##      0      1
## 0.8469 0.1531
##
##
## ## Variable: Complain
```






```
##
## Number of categories: 2
##
## Frequency table
##
##    0    1
## 2011   20
##
## Relative frequency table
##
##      0      1
## 0.9902 0.0098
##
##
## ## Variable: Age
```






```
##
## Summary Statistics for Age
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    24.0    43.0    50.0    51.2    61.0    80.0
##
## Standard deviation: 11.7083
## Coefficient of variation (sd/mean): 0.2287
```

1.7 Bivariate Analysis (when relevant)

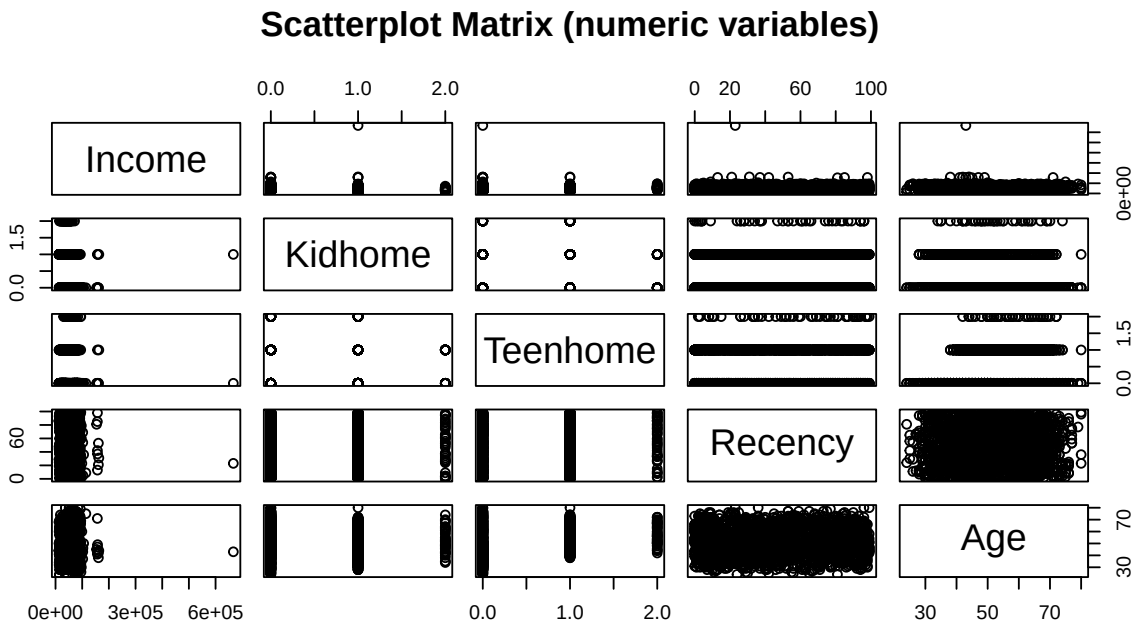
```
# Detect variable types
num_vars <- names(dd)[sapply(dd, is.numeric)]
cat_vars <- names(dd)[sapply(dd, is.factor)]

cat("\n### Numeric-Numeric relationships\n")
```

```
##
## ### Numeric-Numeric relationships
```

```
if (length(num_vars) > 1) {
  cor_mat <- cor(dd[num_vars], use = "pairwise.complete.obs")
  print(round(cor_mat, 3))
  pairs(dd[num_vars], main = "Scatterplot Matrix (numeric variables)")
} else {
  cat("Not enough numeric variables for correlation analysis.\n")
}
```

```
##           Income Kidhome Teenhome Recency    Age
## Income     1.000  -0.427    0.013  -0.009  0.154
## Kidhome   -0.427   1.000   -0.046   0.017 -0.245
## Teenhome   0.013  -0.046    1.000   0.026  0.361
## Recency   -0.009   0.017    0.026   1.000  0.020
## Age        0.154  -0.245    0.361   0.020  1.000
```

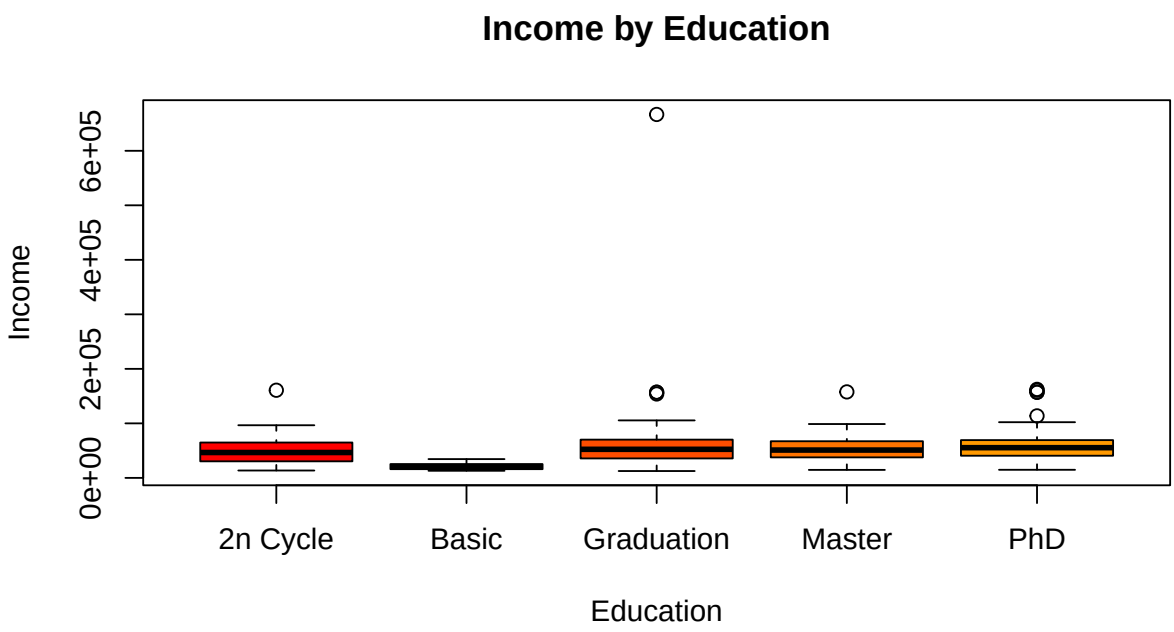


```r
cat("\n### Numeric-Categorical relationships\n")
```

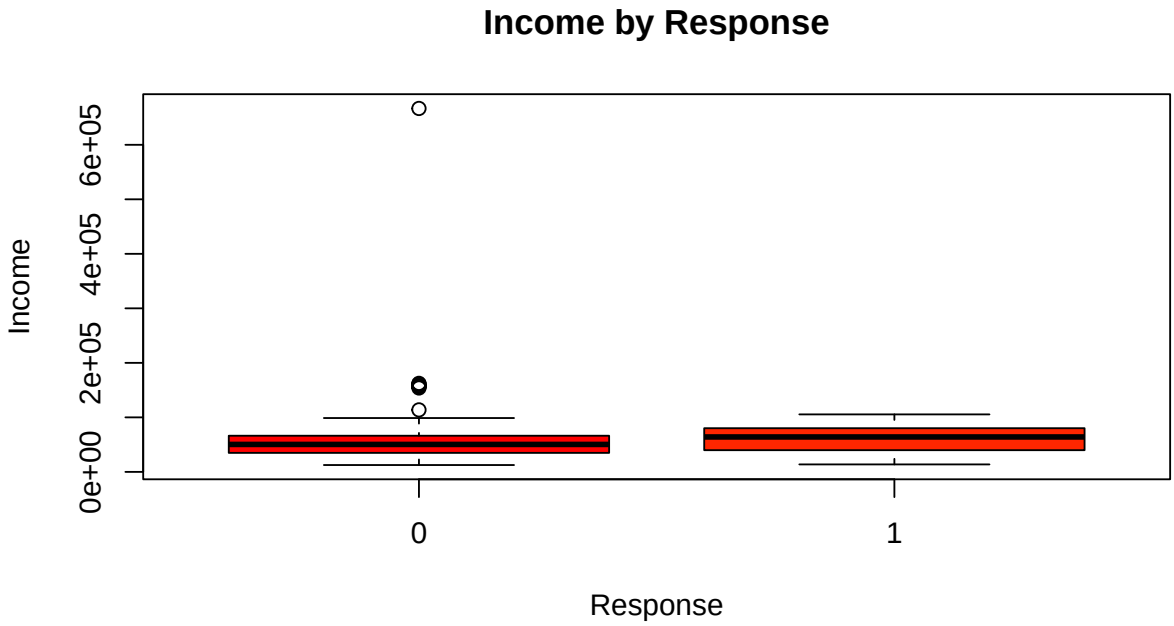
```
##
## ### Numeric-Categorical relationships
```

```r
if (length(num_vars) > 0 & length(cat_vars) > 0) {
  for (num in num_vars) {
    for (catv in cat_vars) {
      cat("\n####", num, "by", catv, "\n")
      boxplot(dd[[num]] ~ dd[[catv]],
              main = paste(num, "by", catv),
              ylab = num, xlab = catv, col = listOfColors)
      if (length(unique(na.omit(dd[[catv]]))) == 2) {
        # t-test if only two groups
        t_res <- t.test(dd[[num]] ~ dd[[catv]])
        print(t_res)
      } else {
        # ANOVA otherwise
        aov_res <- aov(dd[[num]] ~ dd[[catv]])
        print(summary(aov_res))
      }
    }
  }
} else {
  cat("No numeric or categorical variables for mixed analysis.\n")
}
```

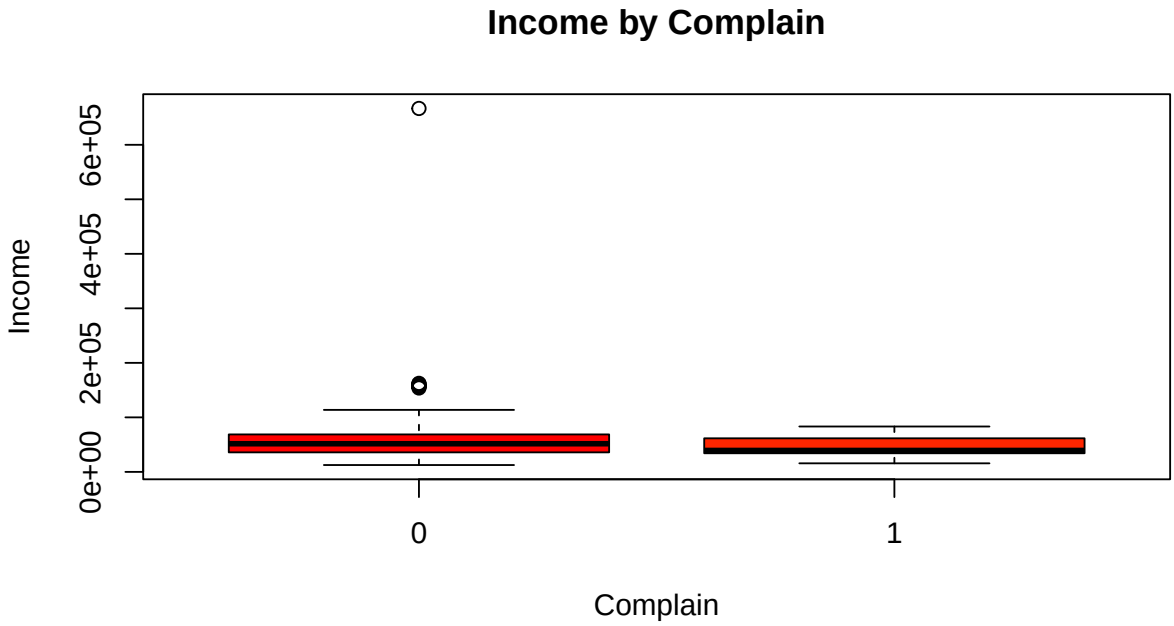
```
##
## #### Income by Education
```



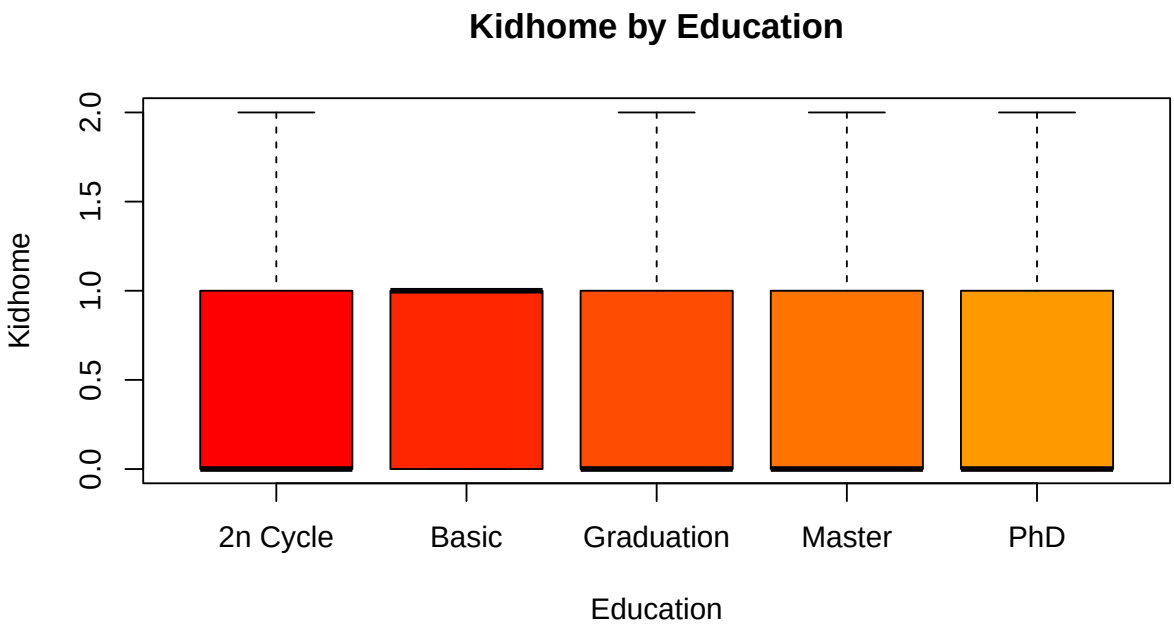

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## dd[[catv]]      4 5.948e+10 1.487e+10   24.41 <2e-16 ***
## Residuals    2026 1.234e+12 6.091e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## #### Income by Response
```




```
##
##  Welch Two Sample t-test
##
## data:  dd[[num]] by dd[[catv]]
## t = -6.1458, df = 459.43, p-value = 1.728e-09
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -11644.525  -6001.958
## sample estimates:
## mean in group 0 mean in group 1
##        51493.36        60316.60
##
##
## #### Income by Complain
```

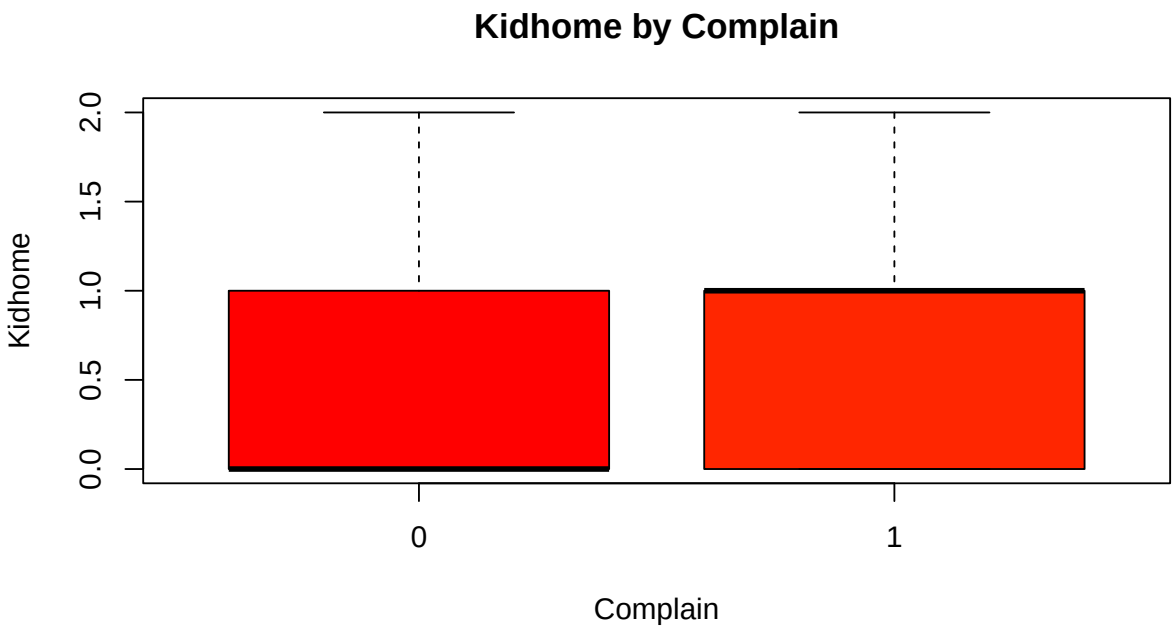



```
##
##  Welch Two Sample t-test
##
## data:  dd[[num]] by dd[[catv]]
## t = 1.6605, df = 19.63, p-value = 0.1127
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -1897.631 16622.503
## sample estimates:
## mean in group 0 mean in group 1
##        52916.94        45554.50
##
##
## #### Kidhome by Education
```

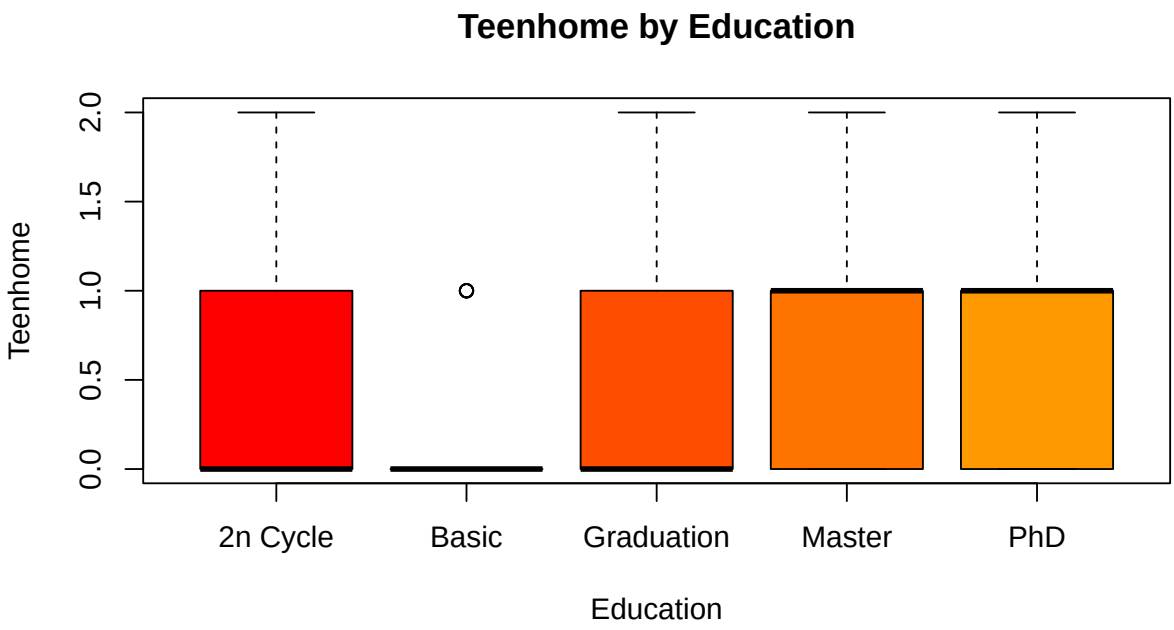



```
##                Df Sum Sq Mean Sq F value Pr(>F)
## dd[[catv]]      4    2.7  0.6864   2.378 0.0498 *
## Residuals    2026  584.8  0.2886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## #### Kidhome by Response
```


```
##
##  Welch Two Sample t-test
##
## data:  dd[[num]] by dd[[catv]]
## t = 3.9996, df = 460.97, p-value = 7.388e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  0.06233116 0.18274100
## sample estimates:
## mean in group 0 mean in group 1
##       0.4633721       0.3408360
##
##
## #### Kidhome by Complain
```

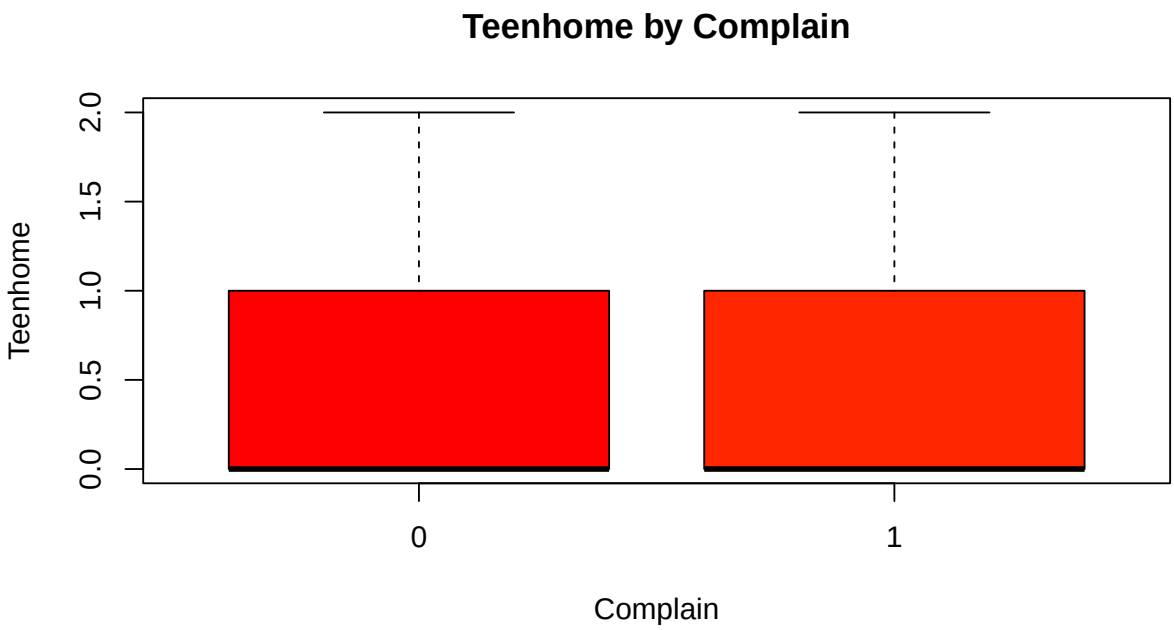



```
##
##  Welch Two Sample t-test
##
## data:  dd[[num]] by dd[[catv]]
## t = -1.5734, df = 19.318, p-value = 0.1319
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -0.48306026  0.06819204
## sample estimates:
## mean in group 0 mean in group 1
##       0.4425659       0.6500000
##
##
## #### Teenhome by Education
```

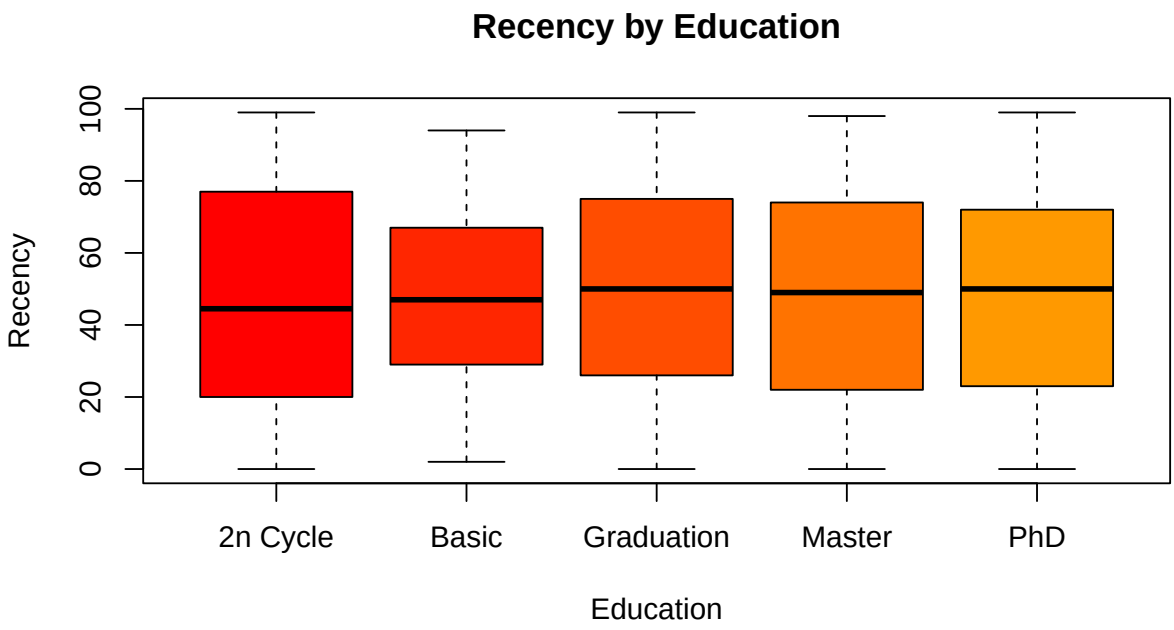



```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## dd[[catv]]      4   14.2   3.559   12.15 9.18e-10 ***
## Residuals    2026  593.4   0.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## #### Teenhome by Response
```

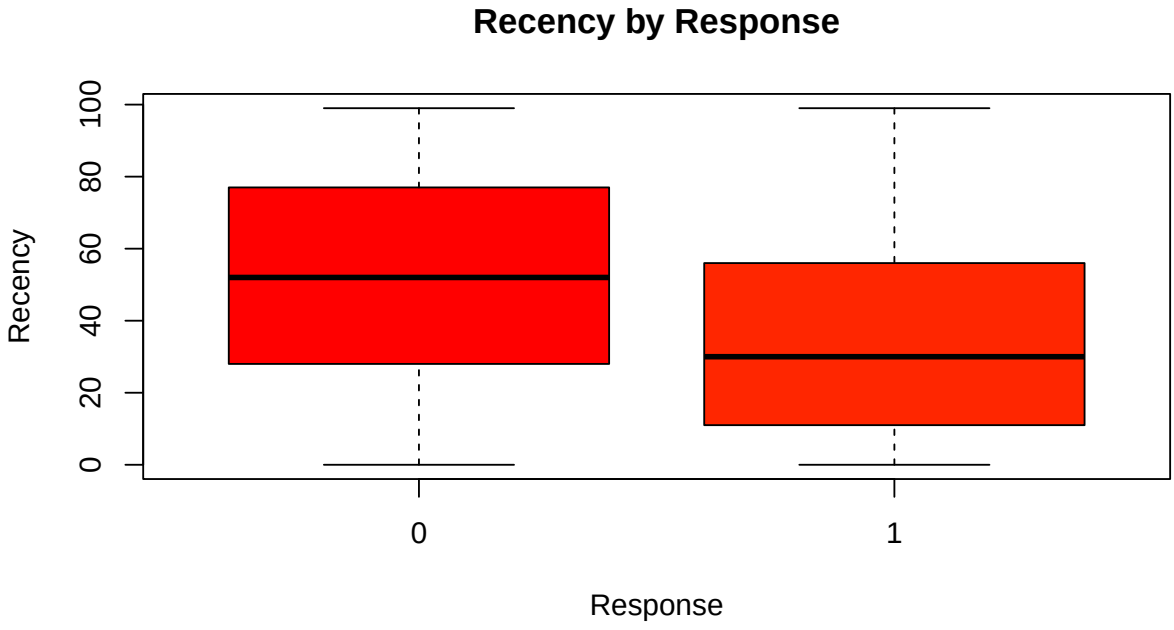

```
##
##  Welch Two Sample t-test
##
## data:  dd[[num]] by dd[[catv]]
## t = 7.731, df = 458.58, p-value = 6.813e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  0.1789067 0.3008585
## sample estimates:
## mean in group 0 mean in group 1
##       0.5453488       0.3054662
##
##
## #### Teenhome by Complain
```



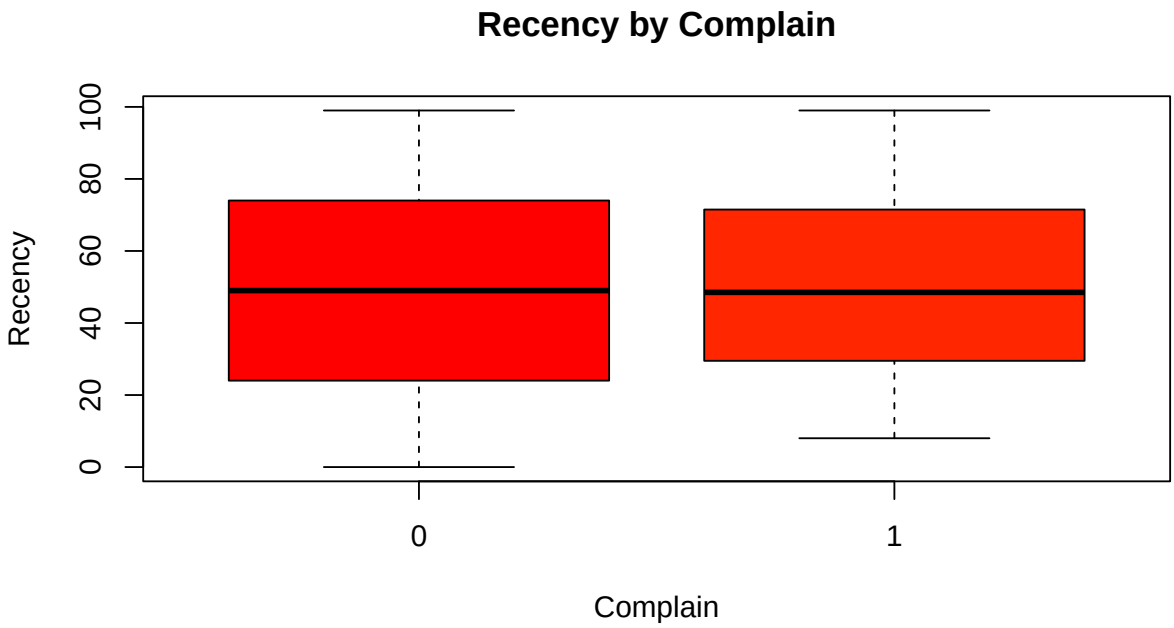

```
##
##  Welch Two Sample t-test
##
## data:  dd[[num]] by dd[[catv]]
## t = 0.063859, df = 19.308, p-value = 0.9497
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -0.2762078  0.2936121
## sample estimates:
## mean in group 0 mean in group 1
##       0.5087021       0.5000000
##
##
## #### Recency by Education
```




```
##                Df  Sum Sq Mean Sq F value Pr(>F)
## dd[[catv]]      4    1760   440.0   0.524  0.718
## Residuals    2026 1702744   840.4
##
## #### Recency by Response
```

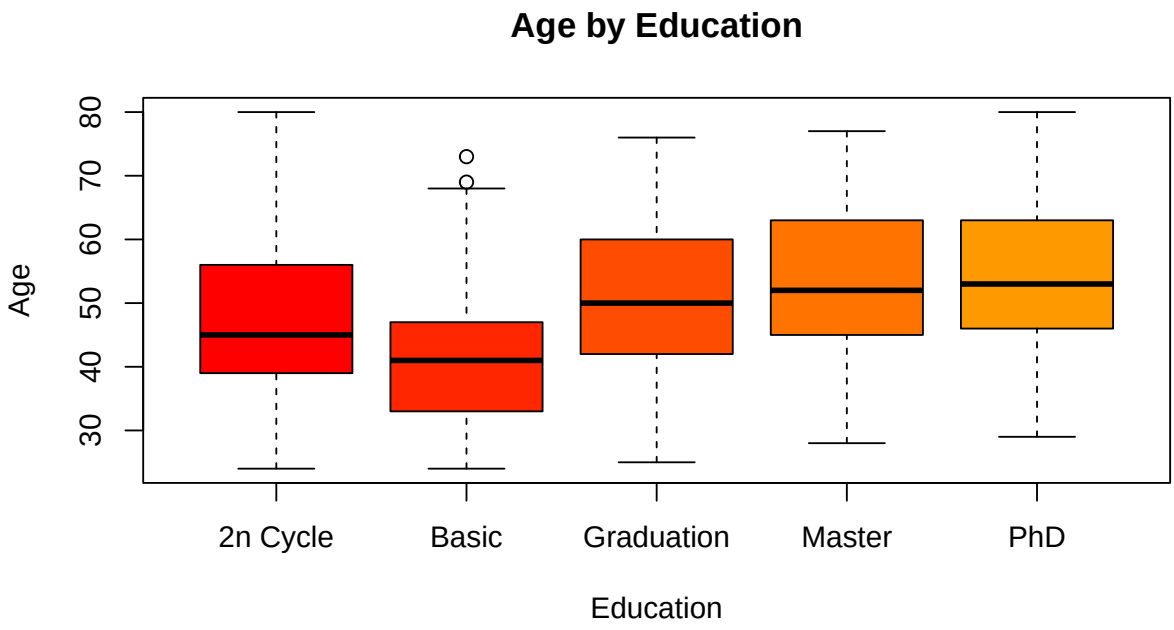


```
##
##  Welch Two Sample t-test
##
## data:  dd[[num]] by dd[[catv]]
## t = 9.7648, df = 440.48, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  13.25037 19.92825
## sample estimates:
## mean in group 0 mean in group 1
##        51.68256        35.09325
##
##
## #### Recency by Complain
```

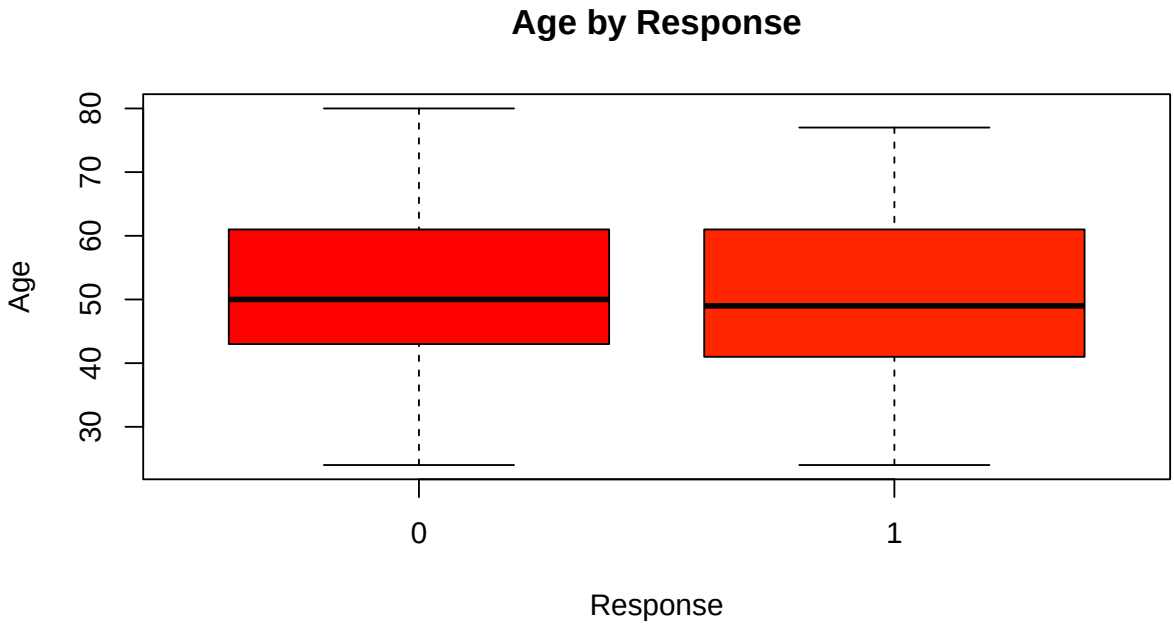



```
##
##  Welch Two Sample t-test
##
## data:  dd[[num]] by dd[[catv]]
## t = -0.31642, df = 19.413, p-value = 0.7551
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -15.03662  11.08227
## sample estimates:
## mean in group 0 mean in group 1
##        49.12282        51.10000
```

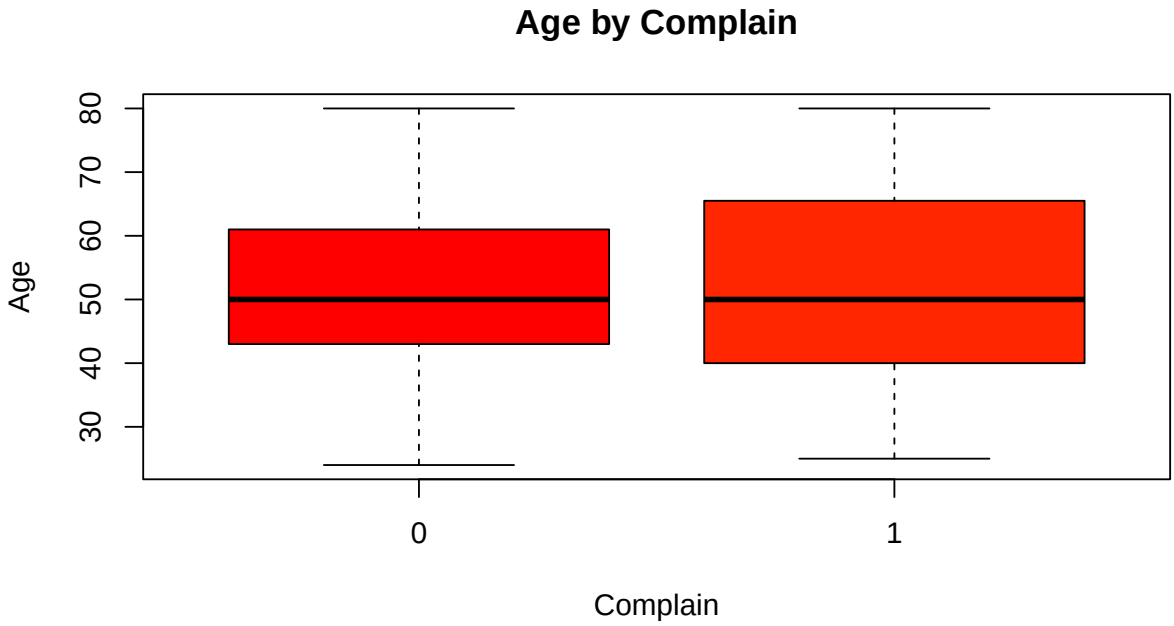
```
##
##
## #### Age by Education
```




```
##                Df Sum Sq Mean Sq F value Pr(>F)
## dd[[catv]]      4  11986  2996.6    22.8 <2e-16 ***
## Residuals    2026 266293   131.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## #### Age by Response
```




```
##
##  Welch Two Sample t-test
##
## data:  dd[[num]] by dd[[catv]]
## t = 0.99319, df = 414.64, p-value = 0.3212
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -0.7333337  2.2311875
## sample estimates:
## mean in group 0 mean in group 1
##        51.31163        50.56270
##
##
## #### Age by Complain
```




```
##
##  Welch Two Sample t-test
##
## data:  dd[[num]] by dd[[catv]]
## t = -0.32382, df = 19.2, p-value = 0.7496
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -8.68602  6.35698
## sample estimates:
## mean in group 0 mean in group 1
##        51.18548        52.35000
```

```
cat("\n### Categorical-Categorical relationships\n")
```

```
##
## ### Categorical-Categorical relationships
```

```
if (length(cat_vars) > 1) {
  for (i in 1:(length(cat_vars)-1)) {
    for (j in (i+1):length(cat_vars)) {
      var1 <- cat_vars[i]; var2 <- cat_vars[j]
      cat("\n####", var1, "vs", var2, "\n")
      tbl <- table(dd[[var1]], dd[[var2]])
      theme_table(as.data.frame.matrix(tbl))
      if (all(dim(tbl) > 1)) {
```

```
        chi <- suppressWarnings(chisq.test(tbl))
        print(chi)
      }
    }
  }
} else {
  cat("Not enough categorical variables for cross-tabulation.\n")
}
```

```
##
## #### Education vs Response
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 19.411, df = 4, p-value = 0.0006524
##
##
## #### Education vs Complain
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 7.0381, df = 4, p-value = 0.1339
##
##
## #### Response vs Complain
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 3.3586e-27, df = 1, p-value = 1
```

1.8 Appendix: Quick numeric summary table

```
num_vars <- names(dd)[sapply(dd, is.numeric)]
if (length(num_vars)) {
  num_summary <- dd %>%
    summarise(across(all_of(num_vars),
                     list(min = ~min(.x, na.rm = TRUE),
                          q1  = ~quantile(.x, 0.25, na.rm = TRUE),
                          mean= ~mean(.x, na.rm = TRUE),
                          median= ~median(.x, na.rm = TRUE),
                          q3  = ~quantile(.x, 0.75, na.rm = TRUE),
                          max = ~max(.x, na.rm = TRUE),
```

```
                          sd  = ~sd(.x, na.rm = TRUE)), .names = "{.col}_{.fn}")) %>%
    pivot_longer(everything(), names_to = c("variable","stat"), names_sep = "_") %>%
    pivot_wider(names_from = stat, values_from = value)
  theme_table(num_summary)
} else {
  cat("No numeric variables detected.")
}
```

variable	min	q1	mean	median	q3	max	sd
Income	12571	35828.5	5.284444e+04	51563	68656	666666	2.524231e+04
Kidhome	0	0.0	4.446086e-01	0	1	2	5.379758e-01
Teenhome	0	0.0	5.086164e-01	0	1	2	5.470923e-01
Recency	0	24.0	4.914229e+01	49	74	99	2.897684e+01
Age	24	43.0	5.119695e+01	50	61	80	1.170826e+01