HW6 - Exploratory Data Analysis (EDA)

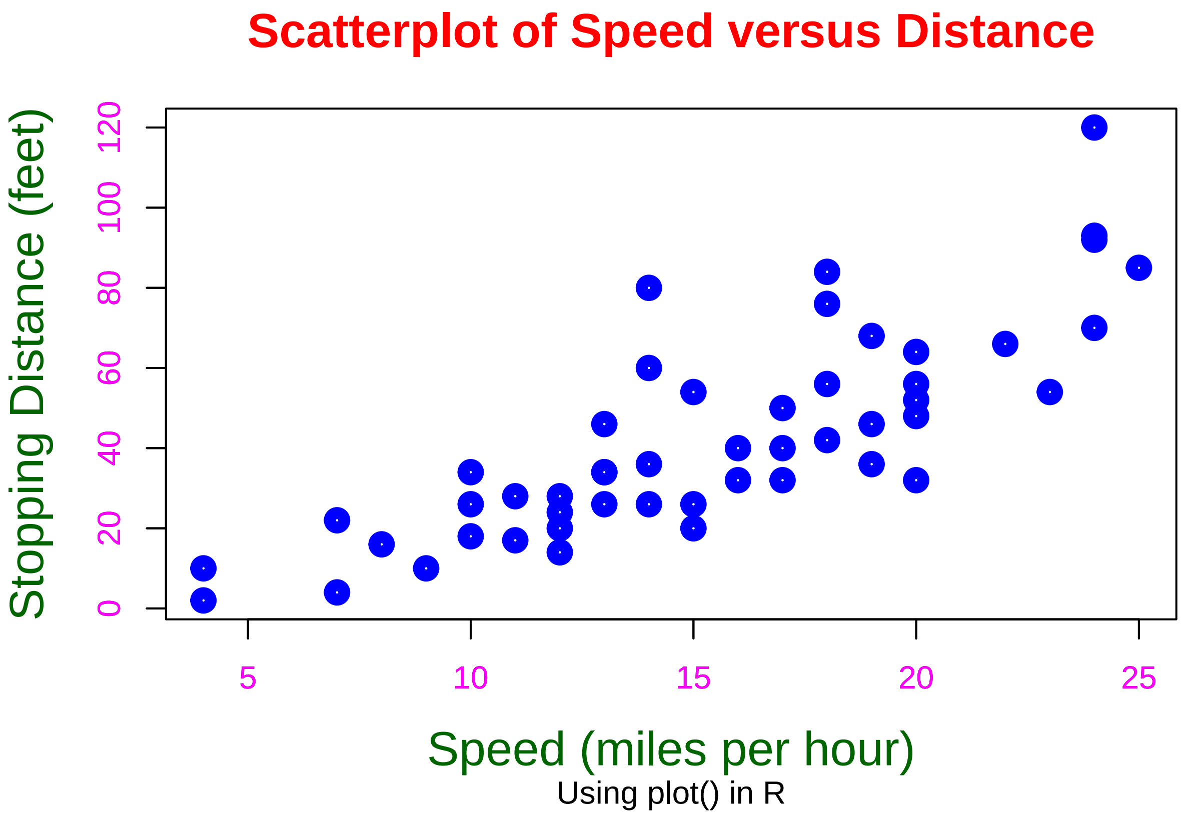
Madhu Peduri

June 27, 2021

0.0.1 1.Use the built-in dataset cars.

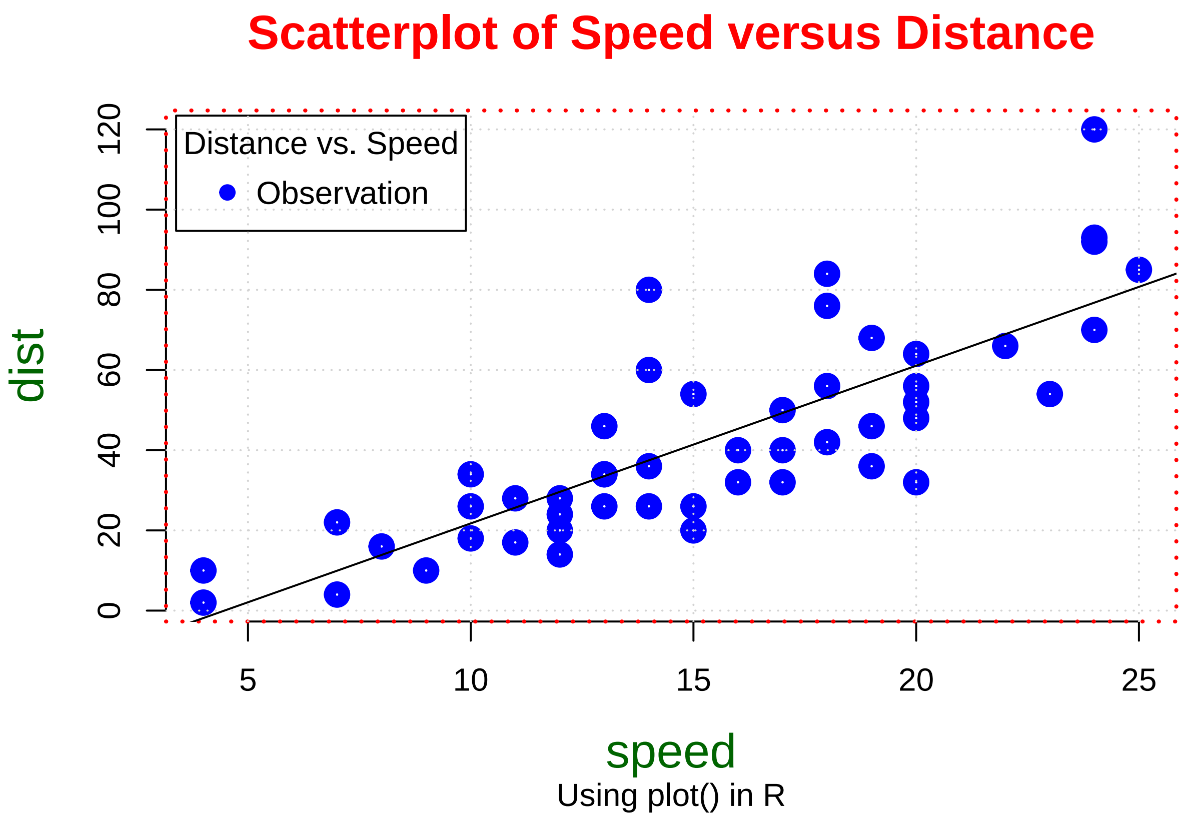
```
car_df <- cars
plot(car_df$speed, car_df$dist, type = "p", col = "blue", lwd = 6,
    xlab = "Speed (miles per hour)", ylab = "Stopping Distance (feet)",
    col.lab = "dark green", cex.lab = 1.5)
title(main = "Scatterplot of Speed versus Distance", cex.main = 1.5,
    col.main = "red", sub = "Using plot() in R", cex.sub = 1)
axis(1, col.axis = "magenta")
axis(2, col.axis = "magenta")
```

0.0.1.1 1.(a) Create a scatter plot of speed versus distance



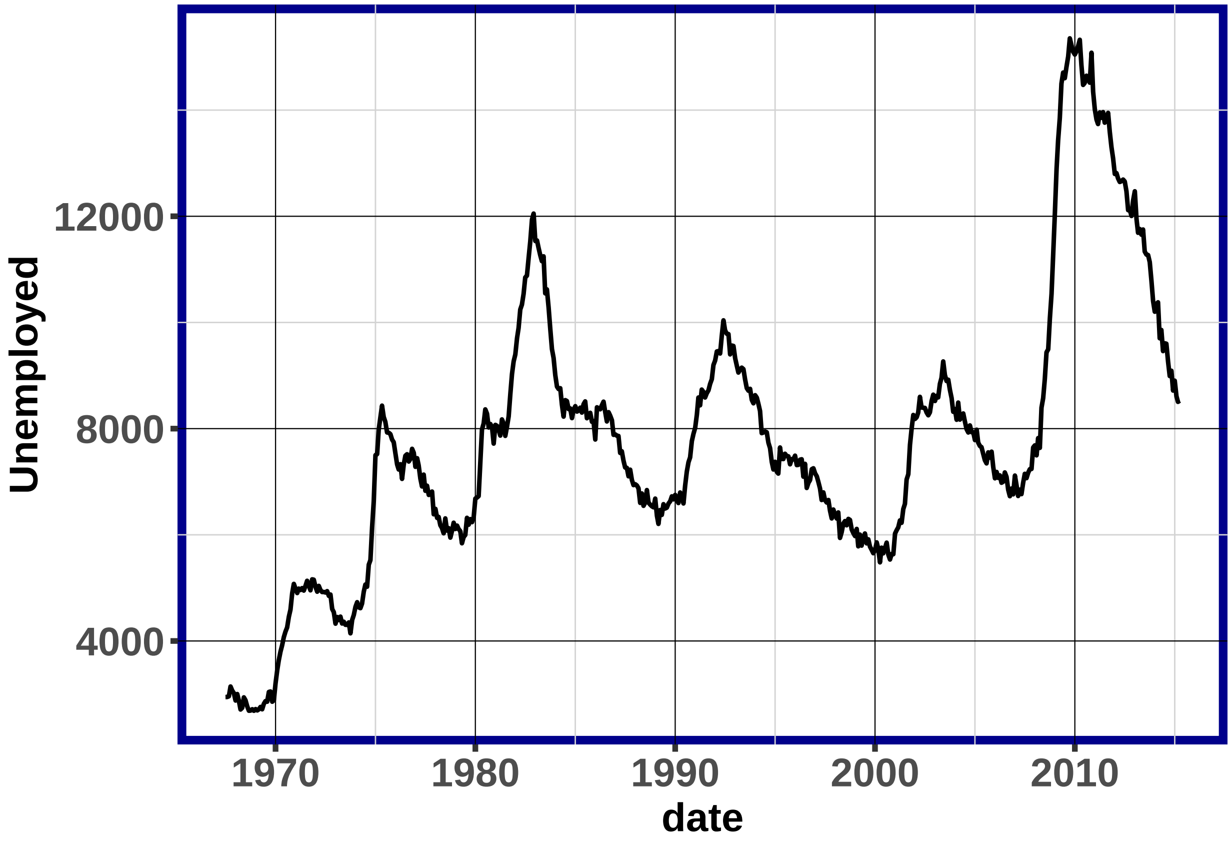
```
car_df <- cars
plot(car_df$speed, car_df$dist, type = "p", col = "blue", lwd = 6,
    xlab = "speed", ylab = "dist", col.lab = "dark green", cex.lab = 1.5,
    axes = FALSE)
title(main = "Scatterplot of Speed versus Distance", cex.main = 1.5,
    col.main = "red", sub = "Using plot() in R", cex.sub = 1)
legend("topleft", inset = 0.01, title = "Distance vs. Speed",
    c("Observation"), col = c("blue"), pch = 19, horiz = TRUE)
axis(1)
axis(2)
box(lty = 3, col = "red", lwd = 2)
grid(lty = 3)
abline(lm(dist ~ speed, data = car_df))
```

0.0.1.2 1.(b) Create Figure 2



0.0.2 2. Use the economics built-in dataset and library ggplot2. Plot the time series of unemployment

```
ggplot(economics, aes(date, unemploy)) + geom_line(size = 0.8) +
    xlab("date") + ylab("Unemployed") + theme(panel.background = element_rect(fill = "white",
    colour = "dark blue", size = 3), panel.grid.major = element_line(colour = "black",
    size = 0.2), panel.grid.minor = element_line(colour = "light gray"),
    axis.ticks = element_line(size = 1), axis.text.x = element_text(size = 15,
        face = "bold"), axis.text.y = element_text(size = 15,
        face = "bold"), axis.title.x = element_text(size = 15,
        face = "bold"), axis.title.y = element_text(size = 15,
        face = "bold"))
```



0.0.3 3. Use the built-in dataset survey

```
str(survey)
```

0.0.3.1 3.(a). Visualize survey dataset

```
## 'data.frame':    237 obs. of  12 variables:
##  $ Sex   : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 2 1 2 2 ...
##  $ Wr.Hnd: num  18.5 19.5 18 18.8 20 18 17.7 17 20 18.5 ...
##  $ NW.Hnd: num  18 20.5 13.3 18.9 20 17.7 17.7 17.3 19.5 18.5 ...
##  $ W.Hnd : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2 2 2 2 2 ...
##  $ Fold  : Factor w/ 3 levels "L on R","Neither",..: 3 3 1 3 2 1 1 3 3 3 ...
##  $ Pulse : int  92 104 87 NA 35 64 83 74 72 90 ...
##  $ Clap  : Factor w/ 3 levels "Left","Neither",..: 1 1 2 2 3 3 3 3 3 3 ...
##  $ Exer  : Factor w/ 3 levels "Freq","None",..: 3 2 2 2 3 3 1 1 3 3 ...
##  $ Smoke : Factor w/ 4 levels "Heavy","Never",..: 2 4 3 2 2 2 2 2 2 2 ...
##  $ Height: num  173 178 NA 160 165 ...
##  $ M.I   : Factor w/ 2 levels "Imperial","Metric": 2 1 NA 2 2 1 1 2 2 2 ...
##  $ Age   : num  18.2 17.6 16.9 20.3 23.7 ...
```

```
describe(survey)
```

```
## survey
##
##  12  Variables      237  Observations
## --------------------------------------------------------------------------------
## Sex
##        n  missing distinct
##      236        1        2
##
## Value      Female   Male
## Frequency     118    118
## Proportion    0.5    0.5
## --------------------------------------------------------------------------------
## Wr.Hnd
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      236        1       60    0.997    18.67     2.09    16.00    16.50
##      .25      .50      .75      .90      .95
##    17.50    18.50    19.80    21.15    22.05
##
## lowest : 13.0 14.0 15.0 15.4 15.5, highest: 22.5 22.8 23.0 23.1 23.2
## --------------------------------------------------------------------------------
## NW.Hnd
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      236        1       68    0.998    18.58    2.184    15.50    16.30
##      .25      .50      .75      .90      .95
##    17.50    18.50    19.72    21.00    22.22
##
## lowest : 12.5 13.0 13.3 13.5 15.0, highest: 22.7 23.0 23.2 23.3 23.5
## --------------------------------------------------------------------------------
## W.Hnd
##        n  missing distinct
##      236        1        2
##
## Value       Left Right
## Frequency     18   218
## Proportion 0.076 0.924
## --------------------------------------------------------------------------------
## Fold
##        n  missing distinct
##      237        0        3
##
## Value       L on R Neither  R on L
## Frequency       99      18     120
## Proportion   0.418   0.076   0.506
## --------------------------------------------------------------------------------
## Pulse
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      192       45       43    0.997    74.15    13.07    59.55    60.00
##      .25      .50      .75      .90      .95
##    66.00    72.50    80.00    90.00    92.00
##
## lowest :  35  40  48  50  54, highest:  96  97  98 100 104
## --------------------------------------------------------------------------------
## Clap
##        n  missing distinct
##      236        1        3
##
## Value         Left Neither   Right
## Frequency       39      50     147
## Proportion   0.165   0.212   0.623
## --------------------------------------------------------------------------------
## Exer
##        n  missing distinct
##      237        0        3
##
## Value       Freq  None  Some
## Frequency    115    24    98
## Proportion 0.485 0.101 0.414
## --------------------------------------------------------------------------------
## Smoke
##        n  missing distinct
##      236        1        4
##
## Value      Heavy Never Occas Regul
## Frequency     11   189    19    17
## Proportion 0.047 0.801 0.081 0.072
## --------------------------------------------------------------------------------
## Height
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      209       28       67    0.999    172.4     11.2    157.0    160.0
##      .25      .50      .75      .90      .95
##    165.0    171.0    180.0    185.4    189.6
##
## lowest : 150 152 152 154 155, highest: 192 193 195 196 200
## --------------------------------------------------------------------------------
## M.I
##        n  missing distinct
##      209       28        2
##
## Value      Imperial   Metric
## Frequency        68      141
## Proportion    0.325    0.675
## --------------------------------------------------------------------------------
## Age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      237        0       88    0.999    20.37    4.353    17.08    17.22
##      .25      .50      .75      .90      .95
##    17.67    18.58    20.17    23.58    30.68
##
## lowest : 16.8 16.9 17.0 17.1 17.2, highest: 41.6 43.8 44.2 70.4 73.0
## --------------------------------------------------------------------------------
```

```r
des(survey)
```

```
##
##  No. of observations =  237
##    Variable      Class           Description
## 1  Sex           factor
## 2  Wr.Hnd        numeric
## 3  NW.Hnd        numeric
## 4  W.Hnd         factor
## 5  Fold          factor
## 6  Pulse         integer
## 7  Clap          factor
## 8  Exer          factor
## 9  Smoke         factor
## 10 Height        numeric
## 11 M.I           factor
## 12 Age           numeric
```

```
summ(survey)
```

```
##
## No. of observations = 237
##
##    Var. name obs. mean   median  s.d.   min.   max.
## 1  Sex       236  1.5    1.5     0.501  1      2
## 2  Wr.Hnd    236  18.67  18.5    1.88   13     23.2
## 3  NW.Hnd    236  18.58  18.5    1.97   12.5   23.5
## 4  W.Hnd     236  1.924  2       0.266  1      2
## 5  Fold      237  2.089  3       0.959  1      3
## 6  Pulse     192  74.15  72.5    11.69  35     104
## 7  Clap      236  2.458  3       0.762  1      3
## 8  Exer      237  1.928  2       0.947  1      3
## 9  Smoke     236  2.178  2       0.621  1      4
## 10 Height    209  172.38 171     9.85   150    200
## 11 M.I       209  1.675  2       0.47   1      2
## 12 Age       237  20.37  18.58   6.47   16.75  73
```

```
codebook(survey)
```

```
##
##
##
## Sex   :
##        Frequency Percent
## Female       118      50
## Male         118      50
##
##  ==================
## Wr.Hnd    :
##  obs. mean   median  s.d.   min.   max.
##  236  18.669 18.5    1.879  13     23.2
##
##  ==================
## NW.Hnd    :
##  obs. mean   median  s.d.   min.   max.
##  236  18.583 18.5    1.967  12.5   23.5
##
##  ==================
## W.Hnd     :
##       Frequency Percent
## Left         18    7.63
## Right       218   92.37
##
```

```
##  ==================
## Fold      :
##         Frequency Percent
## L on R         99   41.77
## Neither        18    7.59
## R on L        120   50.63
##
##  ==================
## Pulse     :
##  obs. mean   median  s.d.   min.   max.
##  192  74.151 72.5    11.687 35     104
##
##  ==================
## Clap      :
##         Frequency Percent
## Left           39    16.5
## Neither        50    21.2
## Right         147    62.3
##
##  ==================
## Exer      :
##      Frequency Percent
## Freq       115    48.5
## None        24    10.1
## Some        98    41.4
##
##  ==================
## Smoke     :
##       Frequency Percent
## Heavy        11    4.66
## Never       189   80.08
## Occas        19    8.05
## Regul        17    7.20
##
##  ==================
## Height    :
##  obs. mean    median  s.d.   min.   max.
##  209  172.381 171     9.848  150    200
##
##  ==================
## M.I   :
##          Frequency Percent
## Imperial        68     32.5
## Metric         141     67.5
##
##  ==================
## Age   :
##  obs. mean   median  s.d.   min.   max.
##  237  20.375 18.583  6.474  16.75  73
##
##  ==================
```

```
summary(survey)
```

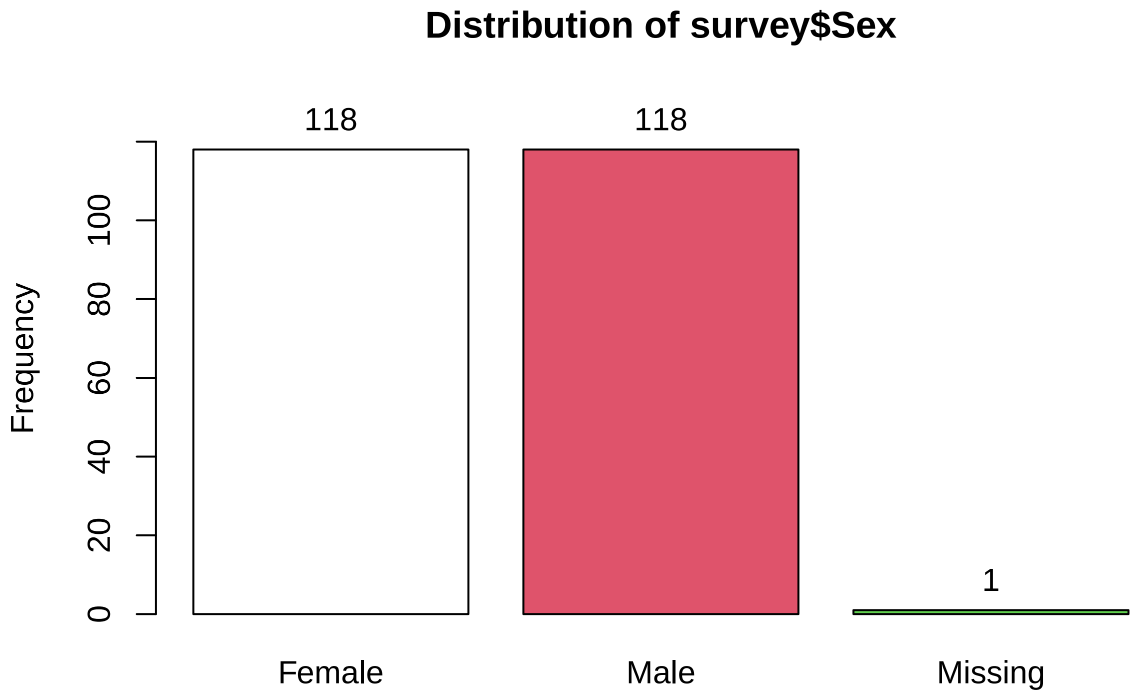
```
##      Sex          Wr.Hnd         NW.Hnd       W.Hnd          Fold
##  Female:118   Min.   :13.0   Min.   :12.5   Left : 18   L on R : 99
##  Male  :118   1st Qu.:17.5   1st Qu.:17.5   Right:218   Neither: 18
##  NA's  :  1   Median :18.5   Median :18.5   NA's :  1   R on L :120
##                Mean   :18.7   Mean   :18.6
##                3rd Qu.:19.8   3rd Qu.:19.7
##                Max.   :23.2   Max.   :23.5
##                NA's   :1      NA's   :1
##      Pulse           Clap       Exer       Smoke         Height
##  Min.   : 35.0   Left   : 39   Freq:115   Heavy: 11   Min.   :150
##  1st Qu.: 66.0   Neither: 50   None: 24   Never:189   1st Qu.:165
##  Median : 72.5   Right  :147   Some: 98   Occas: 19   Median :171
##  Mean   : 74.2   NA's   :  1              Regul: 17   Mean   :172
##  3rd Qu.: 80.0                            NA's :  1   3rd Qu.:180
##  Max.   :104.0                                        Max.   :200
##  NA's   :45                                           NA's   :28
##        M.I          Age
##  Imperial: 68   Min.   :16.8
##  Metric  :141   1st Qu.:17.7
##  NA's    : 28   Median :18.6
##                   Mean   :20.4
##                   3rd Qu.:20.2
##                   Max.   :73.0
##
```

0.0.4 3.(b) Generate frequency for features Sex, Exer and Smoke

```
table(survey$Sex)
```

```
##
## Female   Male
##    118    118
```

```
tab1(survey$Sex)
```

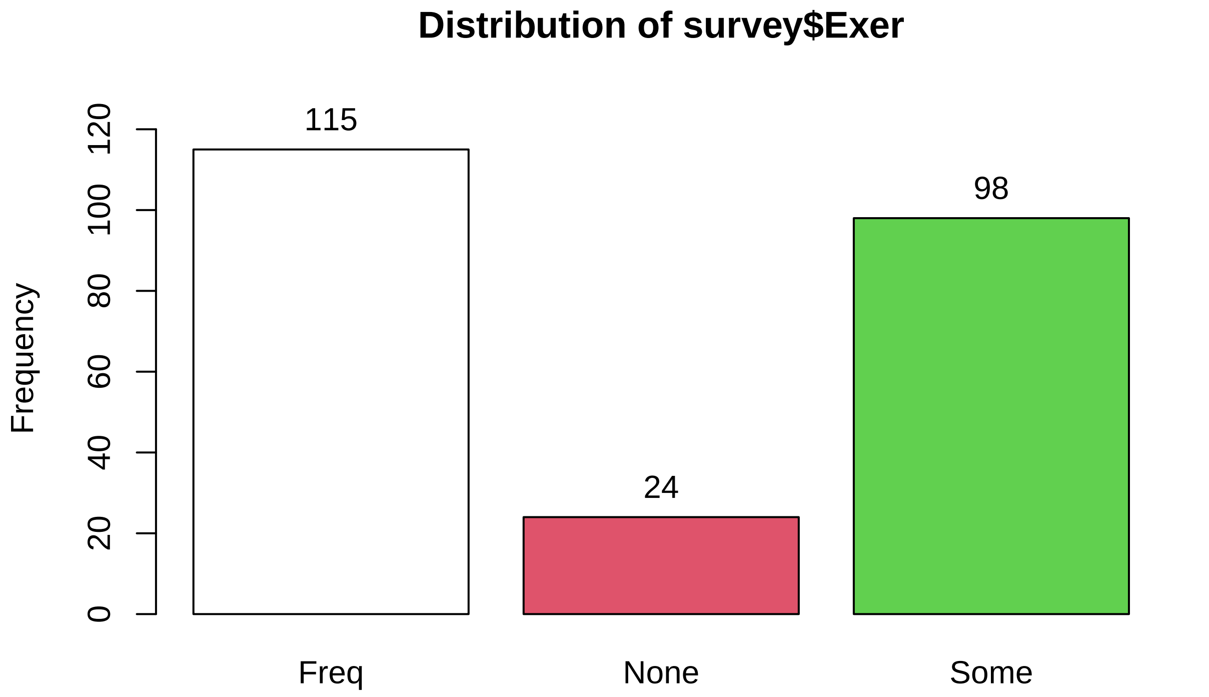


```
## survey$Sex :
##         Frequency   %(NA+)   %(NA-)
## Female        118     49.8       50
## Male          118     49.8       50
## NA's            1      0.4        0
##   Total       237    100.0      100
```

```
table(survey$Exer)
```

```
##
## Freq None Some
##  115   24   98
```

```
tab1(survey$Exer)
```

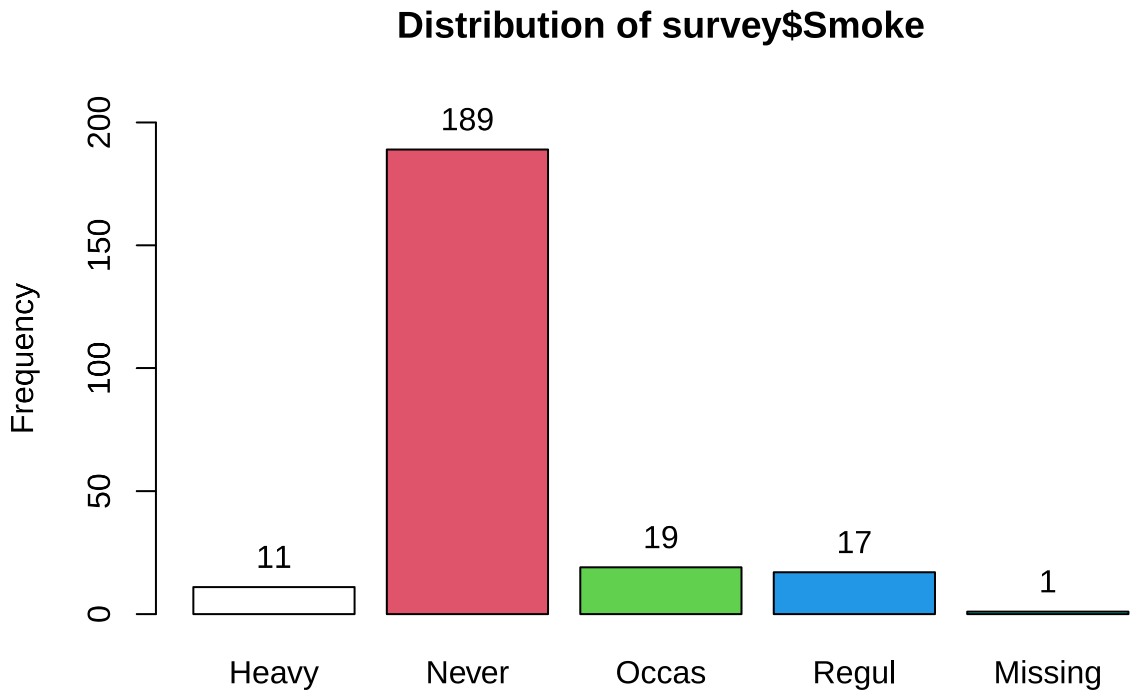


```
## survey$Exer :
##         Frequency Percent Cum. percent
## Freq          115    48.5         48.5
## None           24    10.1         58.6
## Some           98    41.4        100.0
##   Total       237   100.0        100.0
```

```
table(survey$Smoke)
```

```
##
## Heavy Never Occas Regul
##    11   189    19    17
```

```
tab1(survey$Smoke)
```



```
## survey$Smoke :
##          Frequency   %(NA+)   %(NA-)
## Heavy           11      4.6      4.7
## Never          189     79.7     80.1
## Occas           19      8.0      8.1
## Regul           17      7.2      7.2
## NA's             1      0.4      0.0
##   Total        237    100.0    100.0
```

0.0.5 3.(c) Produce contingency tables

```
table(survey$Sex, survey$Exer, useNA = "ifany")
```

```
##
##           Freq None Some
##   Female    49   11   58
##   Male      65   13   40
##   <NA>       1    0    0
```

```
chisq.test(survey$Sex, survey$Exer, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  survey$Sex and survey$Exer
## X-squared = 6, df = 2, p-value = 0.06
```

```
table(survey$Smoke, survey$Exer, useNA = "ifany")
```

```
##
##         Freq None Some
##   Heavy    7    1    3
##   Never   87   18   84
##   Occas   12    3    4
##   Regul    9    1    7
##   <NA>     0    1    0
```

```
chisq.test(survey$Smoke, survey$Exer, correct = FALSE)
```

```
## Warning in chisq.test(survey$Smoke, survey$Exer, correct = FALSE): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  survey$Smoke and survey$Exer
## X-squared = 5, df = 6, p-value = 0.5
```

```
table(survey$Smoke, survey$Sex, useNA = "ifany")
```

```
##
##         Female Male <NA>
##   Heavy      5    6    0
##   Never     99   89    1
##   Occas      9   10    0
##   Regul      5   12    0
##   <NA>       0    1    0
```

```
chisq.test(survey$Smoke, survey$Sex, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  survey$Smoke and survey$Sex
## X-squared = 4, df = 3, p-value = 0.3
```

We can make below observation,

- We have p-value = 0.5 from chi-test between features Somke and Exercies. This is high compared to significant value 0.05 and we can say, there is high correlation between these features.
- We have p-value = 0.3 from chi-test between features Somke and Sex. This is high compared to significant value 0.05 and we can say, there is correlation between these features.
- We have p-value = 0.06 from chi-test between features somke and Sex. This is not high and we can say, there is correlation between these features

0.0.6 3.(d) Correlation matrix

```
data("survey")
ff <- lm(Height ~ Wr.Hnd, data = survey)
summary(ff)
```

```
##
## Call:
## lm(formula = Height ~ Wr.Hnd, data = survey)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.728  -5.071  -0.827   4.947  25.870
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  113.954      5.442    20.9   <2e-16 ***
## Wr.Hnd         3.117      0.289    10.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.91 on 206 degrees of freedom
##   (29 observations deleted due to missingness)
## Multiple R-squared:  0.361,  Adjusted R-squared:  0.358
## F-statistic:  116 on 1 and 206 DF,  p-value: <2e-16
```

```
# calculation of Pearson' correlation coefficient
cor(survey$Wr.Hnd, survey$Height, use = "complete")
```

```
## [1] 0.601
```

```
# This code was used to produce the correlation matrix
library(psych)
```

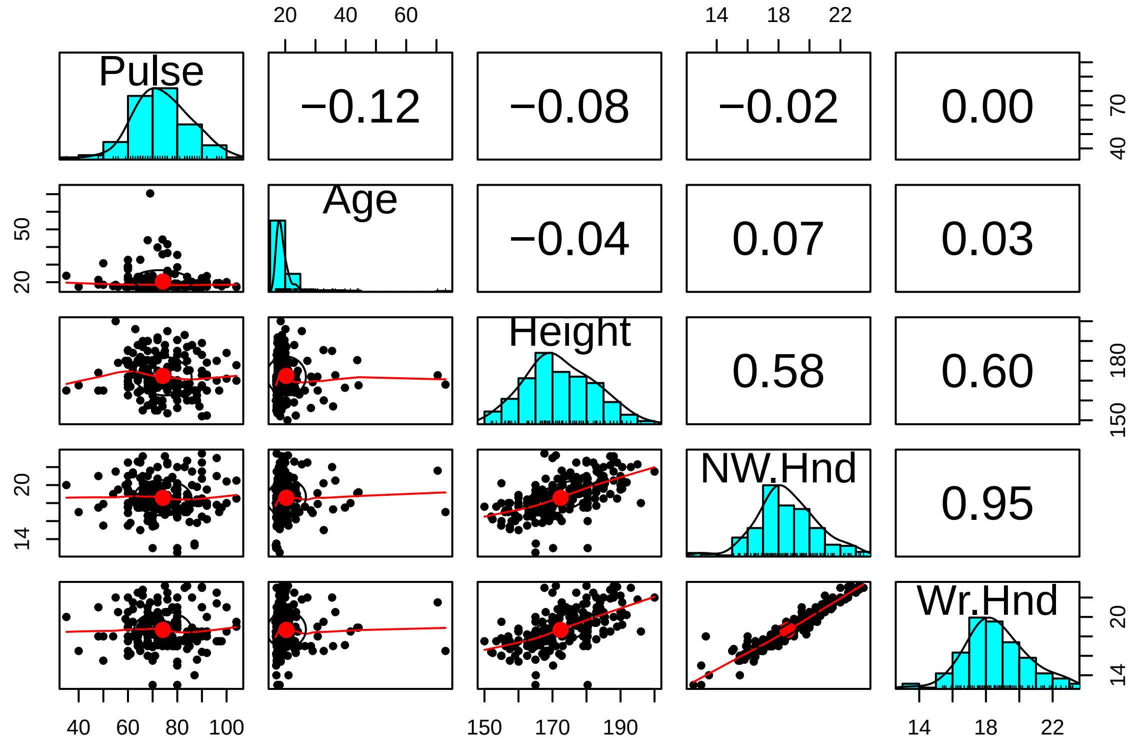
```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:epiDisplay':
##
##     alpha, cs, lookup

## The following object is masked from 'package:Hmisc':
##
##     describe

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
dat0 <- survey[, c("Pulse", "Age", "Height", "NW.Hnd", "Wr.Hnd")]
pairs.panels(dat0)
```



```r
summary(lm(NW.Hnd ~ Wr.Hnd, data = survey))
```

```
##
## Call:
## lm(formula = NW.Hnd ~ Wr.Hnd, data = survey)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.618 -0.404  0.082  0.379  1.683
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0486     0.4075    0.12     0.91
## Wr.Hnd        0.9928     0.0217   45.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.626 on 234 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.899,  Adjusted R-squared:  0.899
## F-statistic: 2.09e+03 on 1 and 234 DF,  p-value: <2e-16
```

We can make below observations,

- Between features, hand span and hight, we have correlation factor as 0.6. By this we can say, there is medium level of correlation, 1 being perfecly correlated.
- We have R-squared = 0.361. From this we can say that the linear regression predictions of the feature

height using hand-span are not good. Low R-square suggest that, there errors are hight between actual and predictions. This is substantiated by the fact that correlation between these two is not high.

- From the Correlation matrix we can say, feature pair ‘Wr.Hnd’ and ‘NW.Hnd’ are hightly correlated with 0.95 value. R-squared for the linear regression between these two features is 0.95 which substantiates the correlation.

0.1 Document Information.

All of the statistical analyses in this document will be performed using R version 4.1.0 (2021-05-18). R packages used will be maintained using the packrat dependency management system.

```
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
## 
## Matrix products: default
## 
## locale:
## [1] LC_COLLATE=English_United States.1252 
## [2] LC_CTYPE=English_United States.1252   
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C                          
## [5] LC_TIME=English_United States.1252    
## 
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods  
## [8] base     
## 
## other attached packages:
##  [1] psych_2.1.6        mnormt_2.0.2       vcd_1.4-8          epiDisplay_3.5.0.1
##  [5] nnet_7.3-16        foreign_0.8-81     Hmisc_4.5-0        Formula_1.2-4     
##  [9] survival_3.2-11    lattice_0.20-44    MASS_7.3-54        ggplot2_3.3.5     
## [13] rmarkdown_2.8      knitr_1.33        
## 
## loaded via a namespace (and not attached):
##  [1] zoo_1.8-9           xfun_0.23           splines_4.1.0      
##  [4] colorspace_2.0-1    vctrs_0.3.8         htmltools_0.5.1.1  
##  [7] yaml_2.2.1          base64enc_0.1-3     utf8_1.2.1         
## [10] rlang_0.4.11        pillar_1.6.1        glue_1.4.2         
## [13] withr_2.4.2         RColorBrewer_1.1-2  jpeg_0.1-8.1       
## [16] lifecycle_1.0.0     stringr_1.4.0       munsell_0.5.0      
## [19] gtable_0.3.0        htmlwidgets_1.5.3   evaluate_0.14      
## [22] labeling_0.4.2      latticeExtra_0.6-29 lmtest_0.9-38      
## [25] parallel_4.1.0      fansi_0.5.0         highr_0.9          
## [28] htmlTable_2.2.1     formatR_1.11        scales_1.1.1       
## [31] backports_1.2.1     checkmate_2.0.0     tmvnsim_1.0-2      
## [34] farver_2.1.0        gridExtra_2.3       png_0.1-7          
## [37] digest_0.6.27       stringi_1.6.1       tools_4.1.0        
## [40] magrittr_2.0.1      tibble_3.1.2        cluster_2.1.2      
## [43] crayon_1.4.1        pkgconfig_2.0.3     ellipsis_0.3.2     
## [46] Matrix_1.3-3        data.table_1.14.0   rstudioapi_0.13    
## [49] R6_2.5.0            rpart_4.1-15        nlme_3.1-152       
## [52] compiler_4.1.0
```