R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Data Introduction

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.3.2
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Load the dataset
df <- read_csv("Placement_Data_Full_Class.csv")
```

```
## Rows: 215 Columns: 15
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): gender, ssc_b, hsc_b, hsc_s, degree_t, workex, specialisation, status
## dbl (7): sl_no, ssc_p, hsc_p, degree_p, etest_p, mba_p, salary
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(df)
```

```
## spc_tbl_ [215 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ sl_no         : num [1:215] 1 2 3 4 5 6 7 8 9 10 ...
##  $ gender        : chr [1:215] "M" "M" "M" "M" ...
##  $ ssc_p         : num [1:215] 67 79.3 65 56 85.8 ...
##  $ ssc_b         : chr [1:215] "Others" "Central" "Central" "Central" ...
##  $ hsc_p         : num [1:215] 91 78.3 68 52 73.6 ...
##  $ hsc_b         : chr [1:215] "Others" "Others" "Central" "Central" ...
##  $ hsc_s         : chr [1:215] "Commerce" "Science" "Arts" "Science" ...
##  $ degree_p      : num [1:215] 58 77.5 64 52 73.3 ...
##  $ degree_t      : chr [1:215] "Sci&Tech" "Sci&Tech" "Comm&Mgmt" "Sci&Tech" ...
```

```
##  $ workex        : chr [1:215] "No" "Yes" "No" "No" ...
##  $ etest_p       : num [1:215] 55 86.5 75 66 96.8 ...
##  $ specialisation: chr [1:215] "Mkt&HR" "Mkt&Fin" "Mkt&Fin" "Mkt&HR" ...
##  $ mba_p         : num [1:215] 58.8 66.3 57.8 59.4 55.5 ...
##  $ status        : chr [1:215] "Placed" "Placed" "Placed" "Not Placed" ...
##  $ salary        : num [1:215] 270000 200000 250000 NA 425000 NA NA 252000 231000 NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   sl_no = col_double(),
##   ..   gender = col_character(),
##   ..   ssc_p = col_double(),
##   ..   ssc_b = col_character(),
##   ..   hsc_p = col_double(),
##   ..   hsc_b = col_character(),
##   ..   hsc_s = col_character(),
##   ..   degree_p = col_double(),
##   ..   degree_t = col_character(),
##   ..   workex = col_character(),
##   ..   etest_p = col_double(),
##   ..   specialisation = col_character(),
##   ..   mba_p = col_double(),
##   ..   status = col_character(),
##   ..   salary = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(df)
```

```
## # A tibble: 6 x 15
##   sl_no gender ssc_p ssc_b   hsc_p hsc_b  hsc_s degree_p degree_t workex etest_p
##   <dbl> <chr>  <dbl> <chr>   <dbl> <chr>  <chr>    <dbl> <chr>    <chr>    <dbl>
## 1     1 M       67   Others   91   Others Comm~     58   Sci&Tech No        55
## 2     2 M       79.3 Central  78.3 Others Scie~     77.5 Sci&Tech Yes       86.5
## 3     3 M       65   Central  68   Centr~ Arts      64   Comm&Mg~ No        75
## 4     4 M       56   Central  52   Centr~ Scie~     52   Sci&Tech No        66
## 5     5 M       85.8 Central  73.6 Centr~ Comm~     73.3 Comm&Mg~ No        96.8
## 6     6 M       55   Others   49.8 Others Scie~     67.2 Sci&Tech Yes       55
## # i 4 more variables: specialisation <chr>, mba_p <dbl>, status <chr>,
## #   salary <dbl>
```

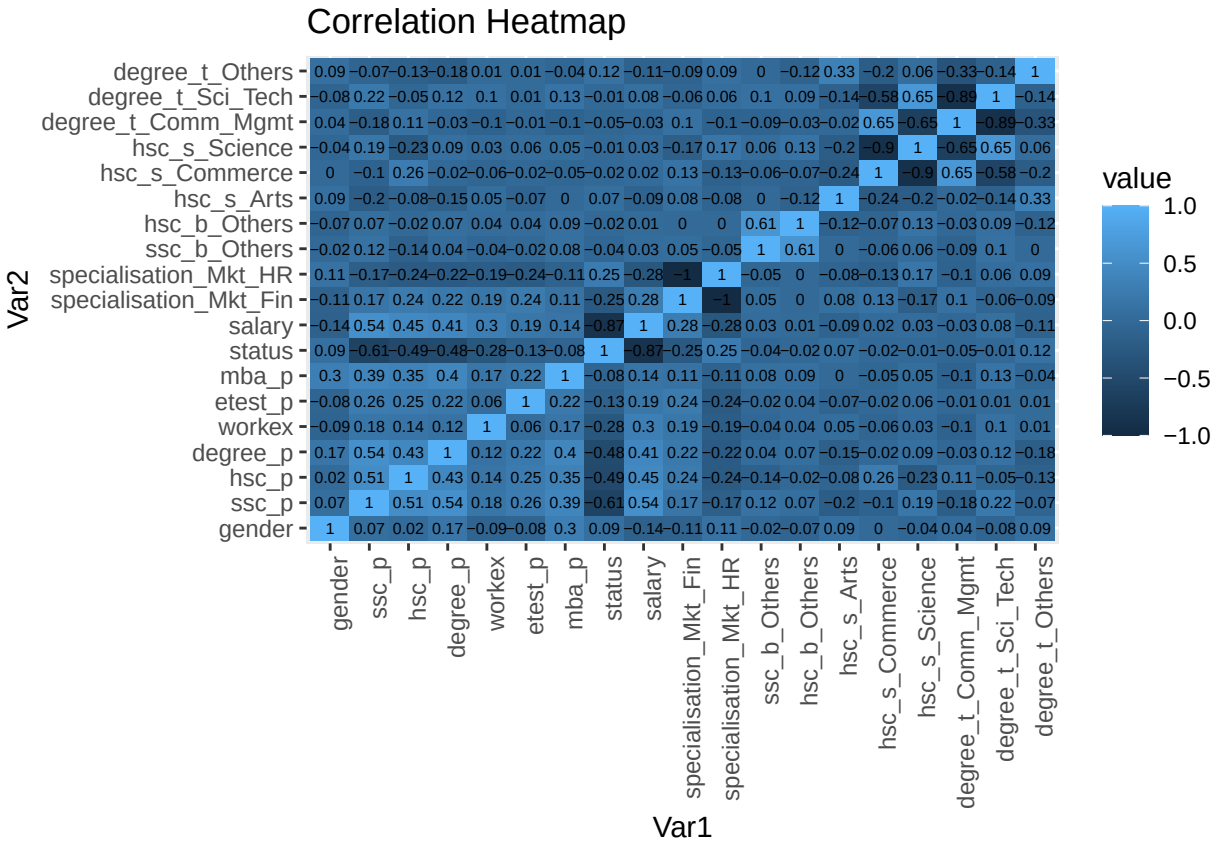
Data Cleaning

```
# Removing unnecessary columns
college_df <- df %>% select(-sl_no)

# Cleaning up null data
college_df$salary[is.na(college_df$salary)] <- 0

# One Hot Encoding
college_df <- college_df %>%
  mutate(specialisation_Mkt_Fin = as.integer(specialisation == 'Mkt&Fin'),
         specialisation_Mkt_HR = as.integer(specialisation == 'Mkt&HR'),
         ssc_b_Others = as.integer(ssc_b == 'Others'),
         hsc_b_Others = as.integer(hsc_b == 'Others'),
```

```
         hsc_s_Arts = as.integer(hsc_s == 'Arts'),
         hsc_s_Commerce = as.integer(hsc_s == 'Commerce'),
         hsc_s_Science = as.integer(hsc_s == 'Science'),
         degree_t_Comm_Mgmt = as.integer(degree_t == 'Comm&Mgmt'),
         degree_t_Sci_Tech = as.integer(degree_t == 'Sci&Tech'),
         degree_t_Others = as.integer(degree_t == 'Others'))

college_df <- college_df %>% select(-c(specialisation, ssc_b, hsc_b, hsc_s, degree_t))

# Dictionary replacement
college_df$gender <- ifelse(college_df$gender == 'M', 0, 1)
college_df$workex <- ifelse(college_df$workex == 'No', 0, 1)
college_df$status <- ifelse(college_df$status == 'Placed', 0, 1)
```

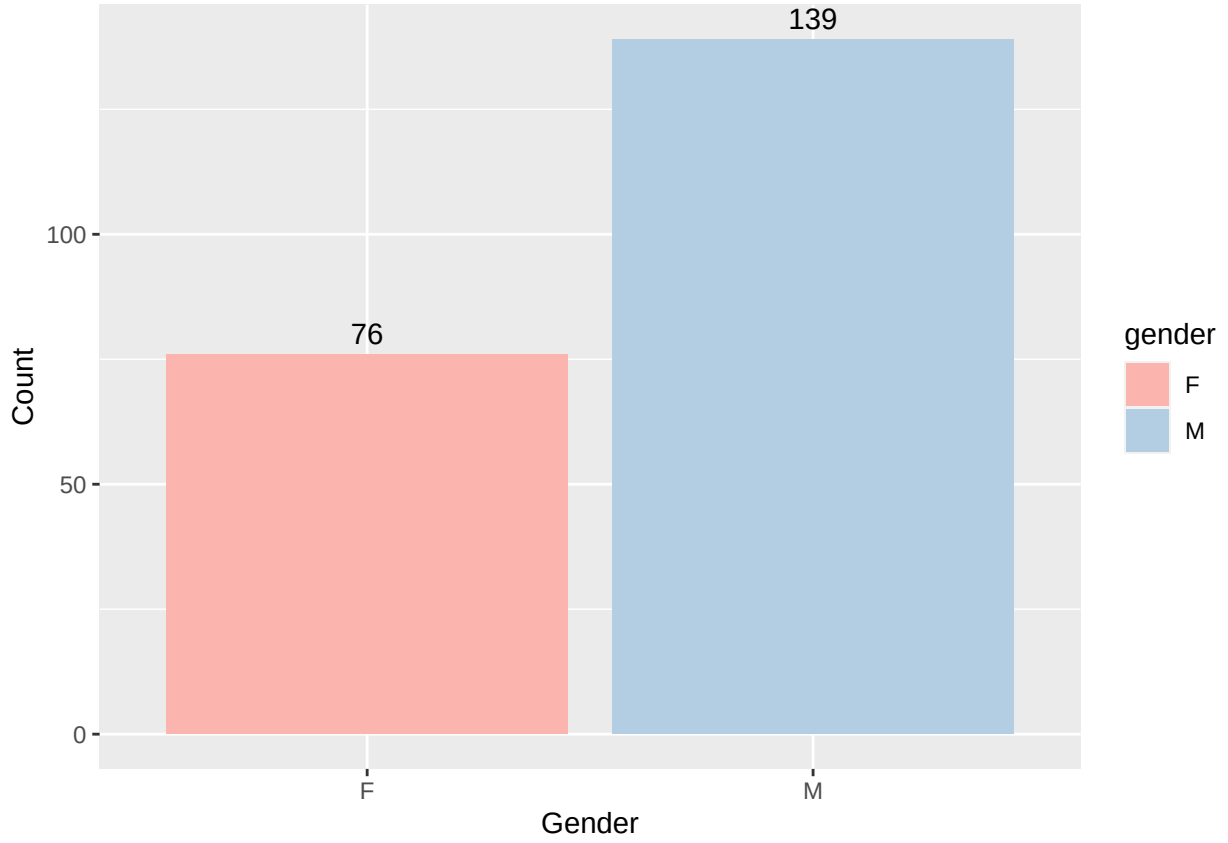
Data Visualization

```
# Load necessary packages
library(reshape2)
library(ggplot2)

# Create a correlation matrix
corr_mat <- round(cor(college_df), 2)

# Melt the correlation matrix for plotting
melted_corr_mat <- melt(corr_mat)

ggplot(data = melted_corr_mat, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(width = 1, height = 1) +
  geom_text(aes(label = value), size = 2) +
  labs(title = "Correlation Heatmap") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

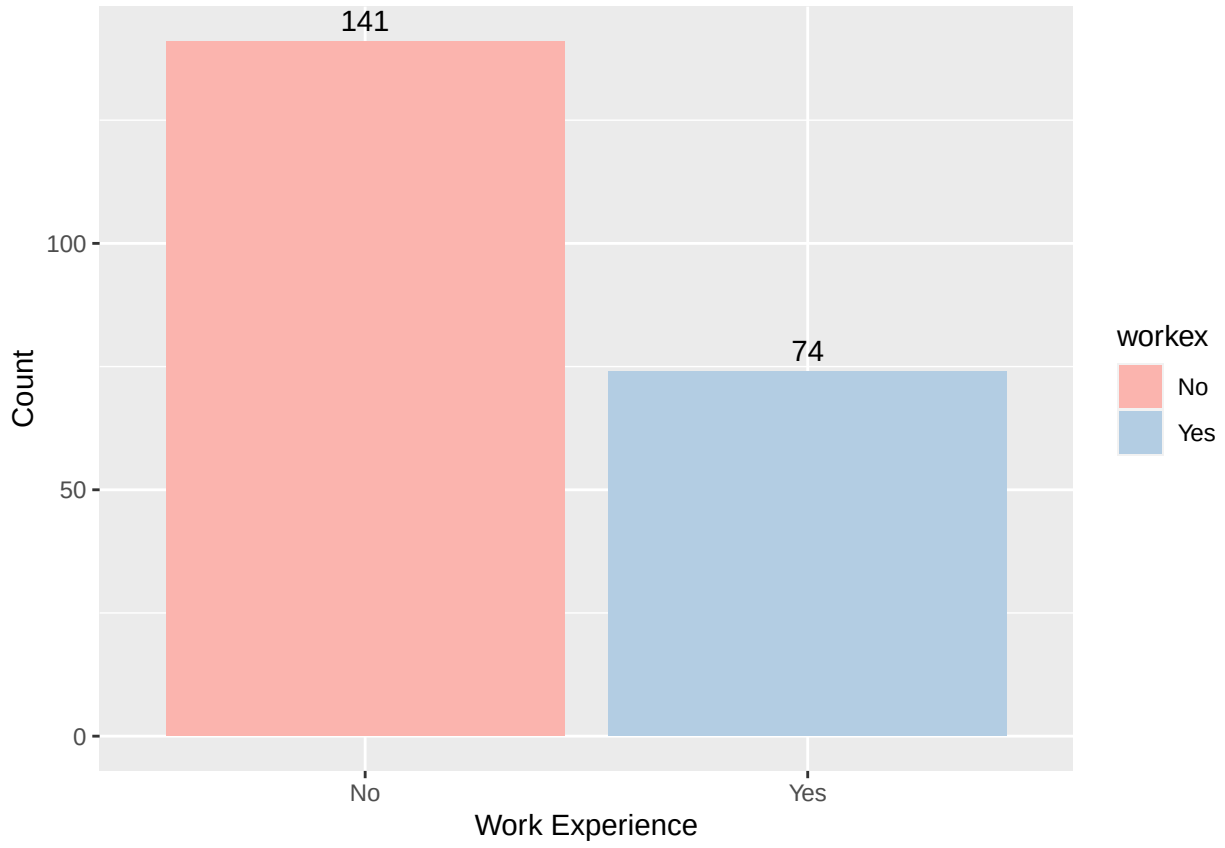


Count Plots

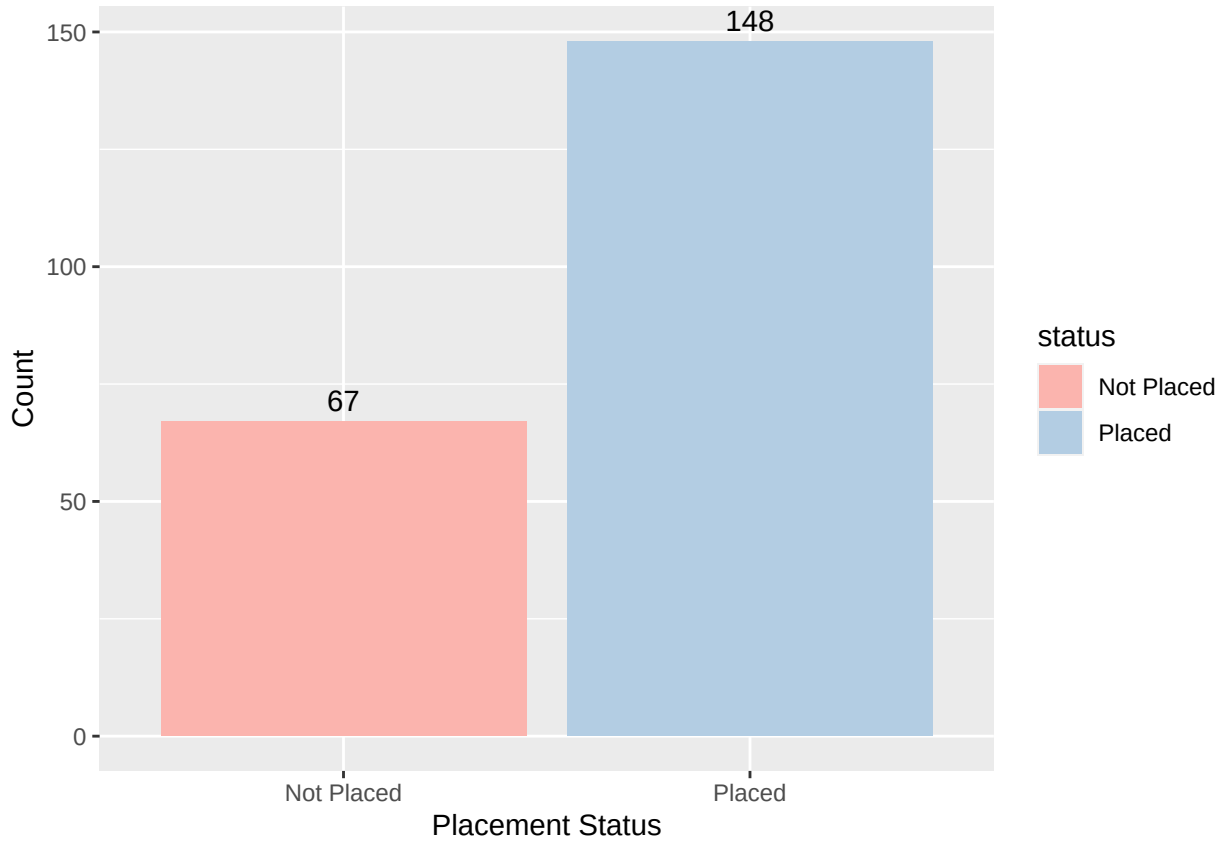
```
# Gender Count
ggplot(df, aes(x = gender, fill = gender)) +
  geom_bar() +
  geom_text(stat = 'count', aes(label = after_stat(count)), vjust = -0.5) +
  labs(x = 'Gender', y = 'Count') +
  scale_fill_brewer(palette = 'Pastel1')
```



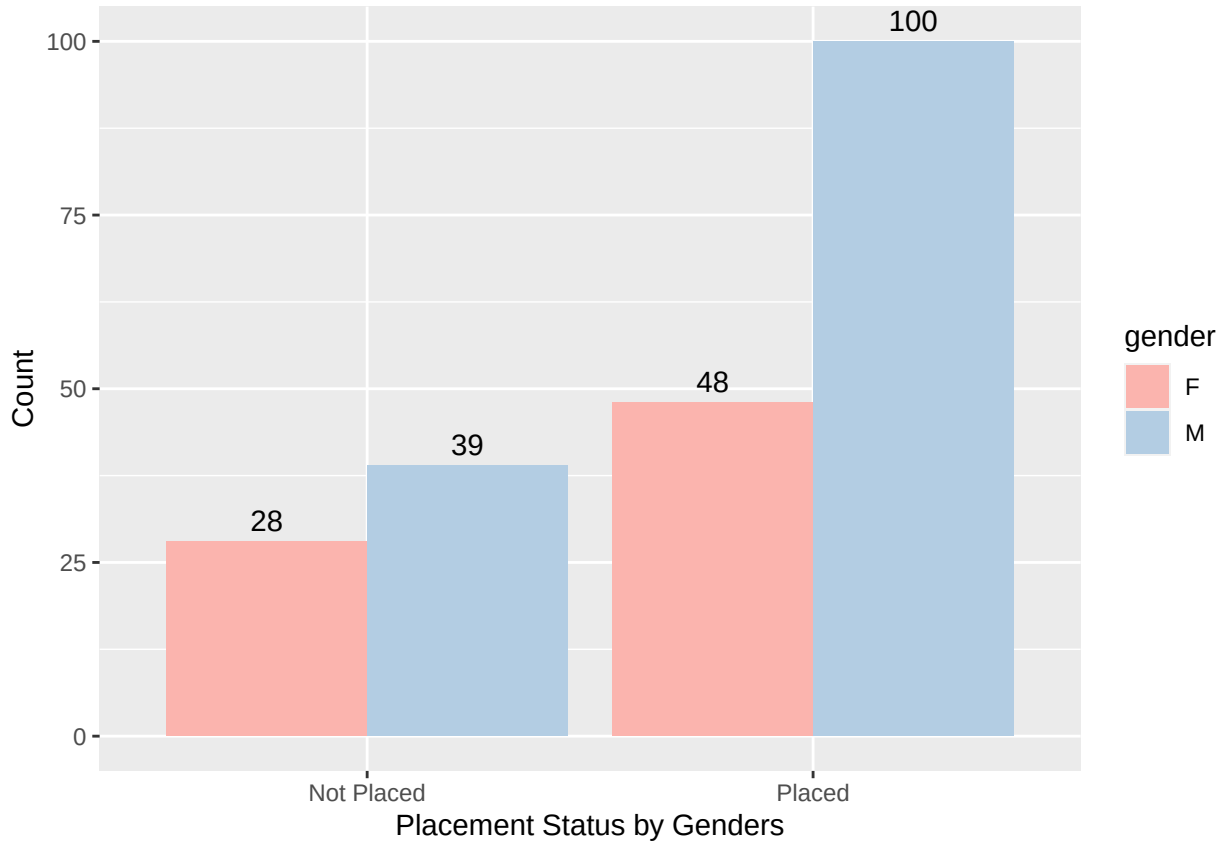

```
# Work Experience
ggplot(df, aes(x = workex, fill = workex)) +
  geom_bar() +
  geom_text(stat = 'count', aes(label = after_stat(count)), vjust = -0.5) +
  labs(x = 'Work Experience', y = 'Count') +
  scale_fill_brewer(palette = 'Pastel1')
```




```
# Placement Status
ggplot(df, aes(x = status, fill = status)) +
  geom_bar() +
  geom_text(stat = 'count', aes(label = after_stat(count)), vjust = -0.5) +
  labs(x = 'Placement Status', y = 'Count') +
  scale_fill_brewer(palette = 'Pastel1')
```

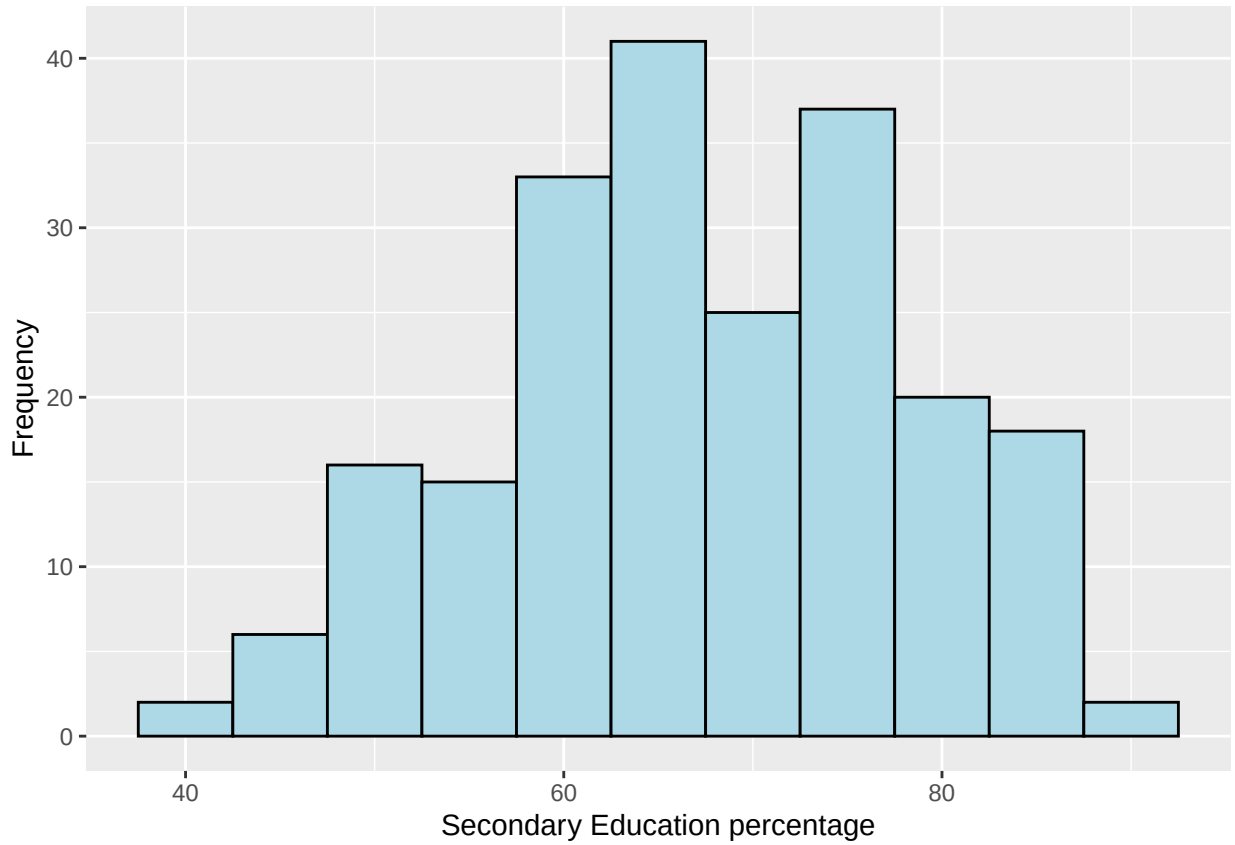



```
# Placement Status by Genders
ggplot(df, aes(x = status, fill = gender)) +
  geom_bar(position = 'dodge') +
  geom_text(stat = 'count', aes(label = after_stat(count)), position = position_dodge(0.9), vjust = -0.5
  labs(x = 'Placement Status by Genders', y = 'Count') +
  scale_fill_brewer(palette = 'Pastel1')
```

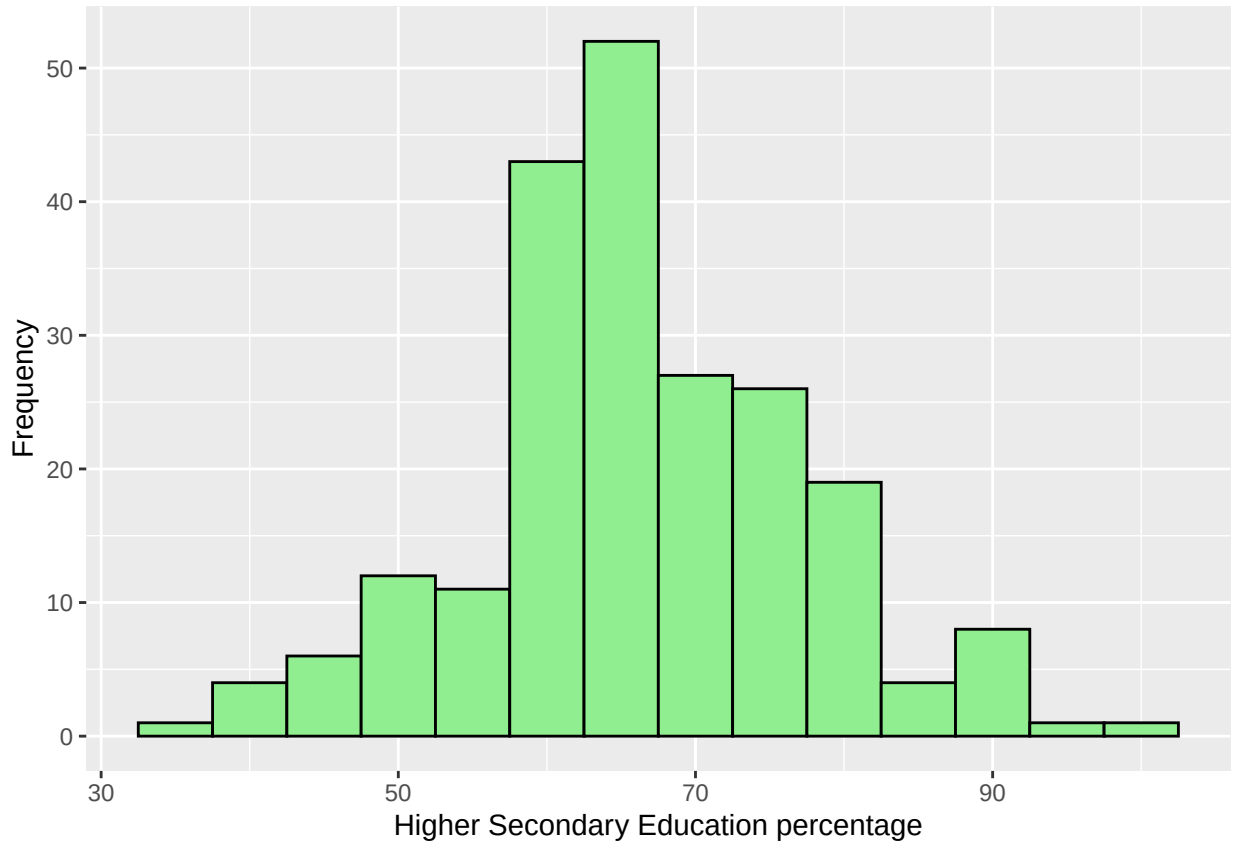


Distribution Plots

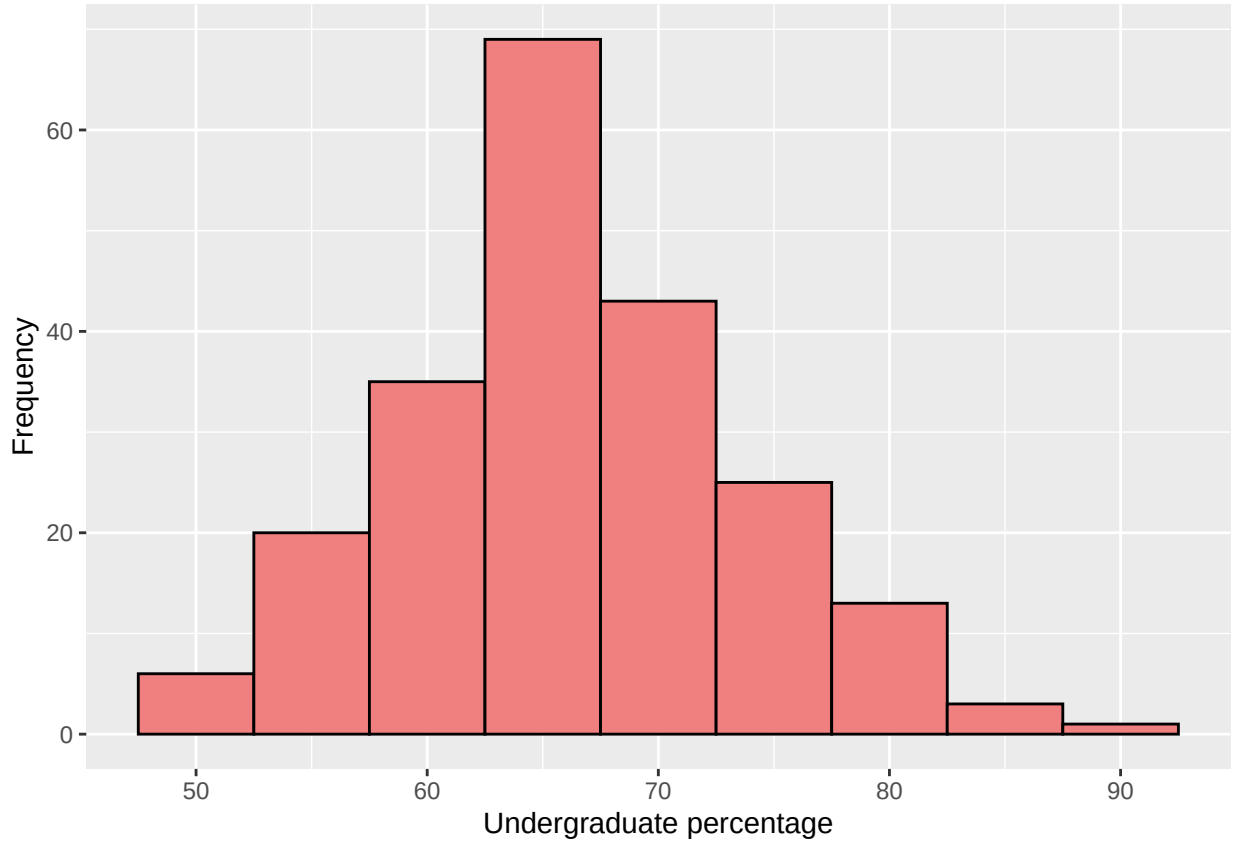
```
# Set the plot size
options(repr.plot.width = 15, repr.plot.height = 8)

# Create distribution plots
par(mfrow = c(2, 2))

# Secondary Education percentage
ggplot(df, aes(x = ssc_p)) +
  geom_histogram(binwidth = 5, fill = 'lightblue', color = 'black') +
  labs(x = 'Secondary Education percentage', y = 'Frequency')
```



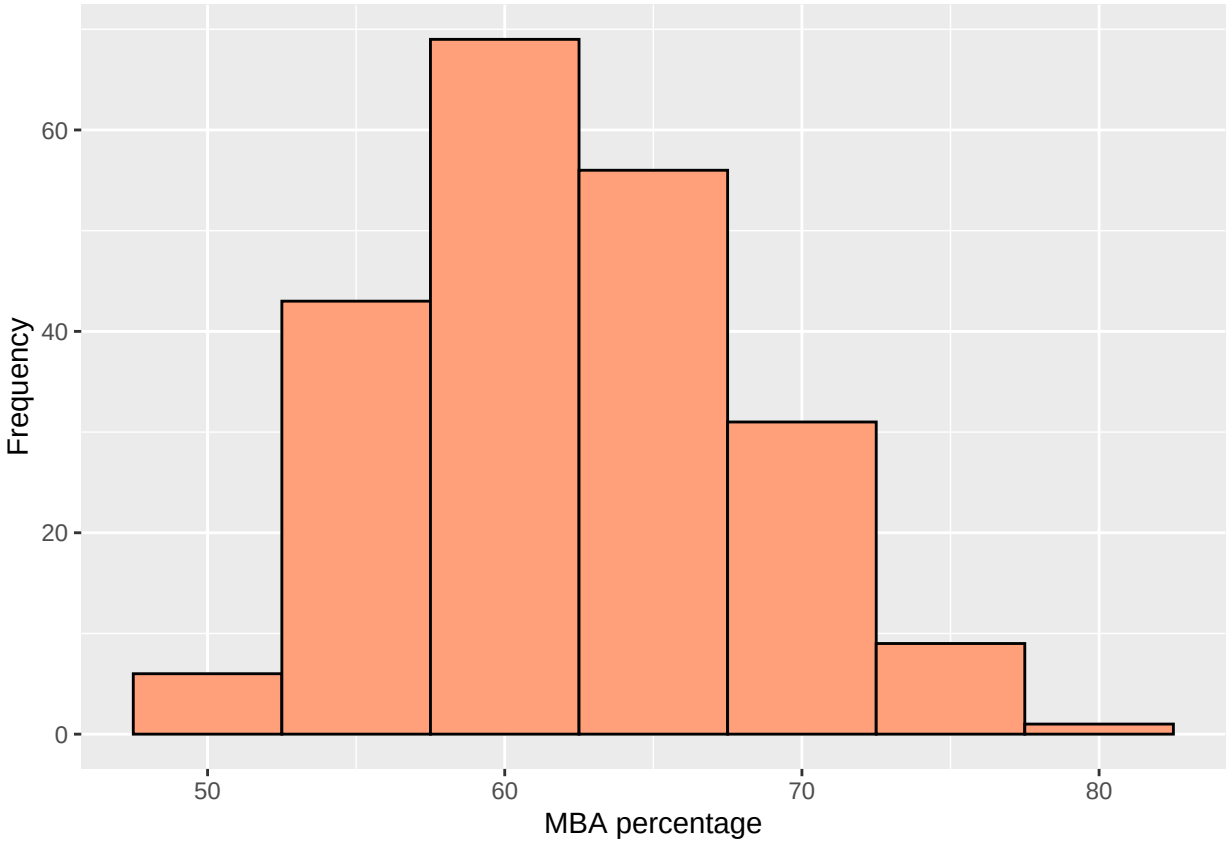

```
# Higher Secondary Education percentage
ggplot(df, aes(x = hsc_p)) +
  geom_histogram(binwidth = 5, fill = 'lightgreen', color = 'black') +
  labs(x = 'Higher Secondary Education percentage', y = 'Frequency')
```




```
# Undergraduate percentage
ggplot(df, aes(x = degree_p)) +
  geom_histogram(binwidth = 5, fill = 'lightcoral', color = 'black') +
  labs(x = 'Undergraduate percentage', y = 'Frequency')
```




```
# MBA percentage
ggplot(df, aes(x = mba_p)) +
  geom_histogram(binwidth = 5, fill = 'lightsalmon', color = 'black') +
  labs(x = 'MBA percentage', y = 'Frequency')
```



Random Forest

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.