# STATS402 final project EDA

Elbert Liu

2024-12-02

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.2
```

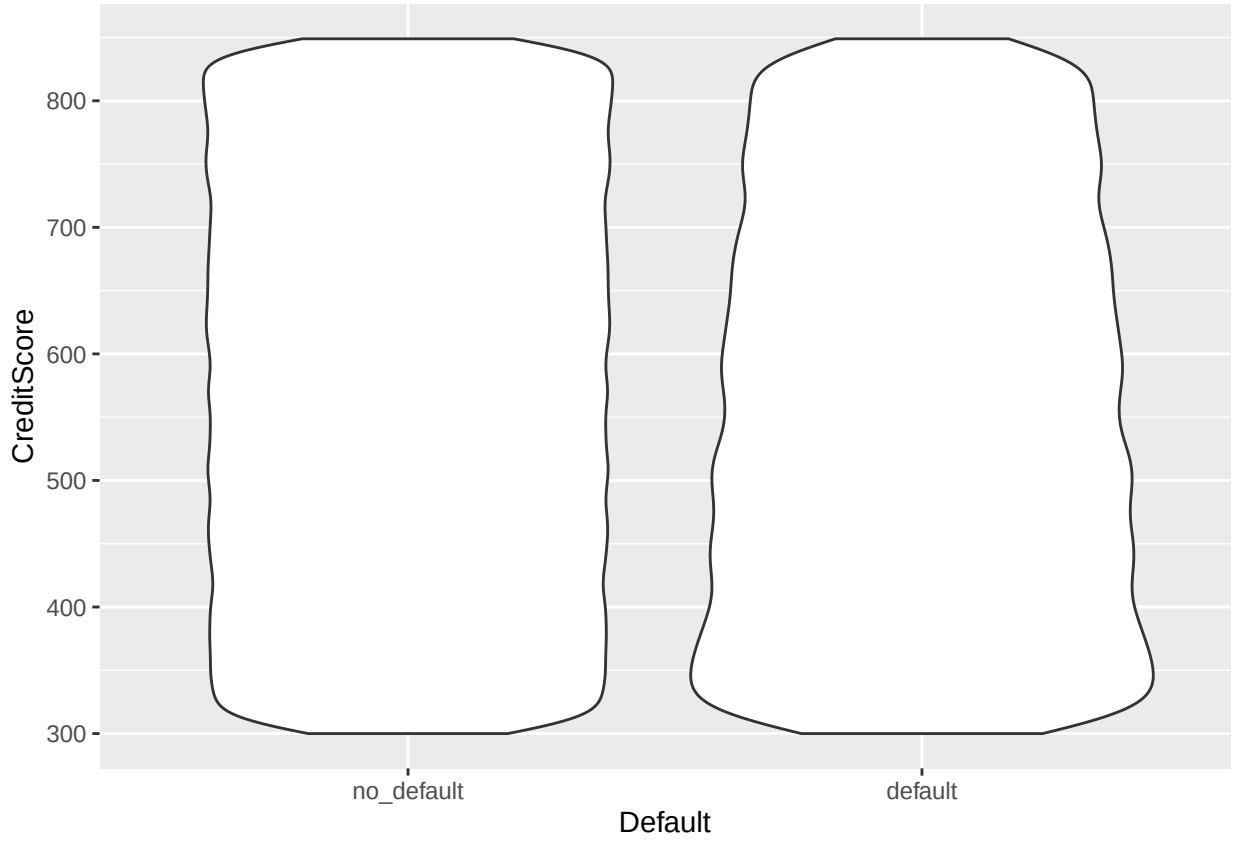
```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```r
library(ggplot2)
df <- read.csv("Loan_default.csv", stringsAsFactors = T)
df <- df[,-1] #Drop ID as it is not necessary
df <- unique(df)
df$Default[df$Default == 0] <- 'no_default'
df$Default[df$Default == 1] <- 'default'
df$Default <- factor(df$Default,levels = c('no_default','default'))
df$Education <- factor(df$Education,levels = c("High School","Bachelor's","Master's","PhD")) #Reorder L
df$MaritalStatus <- factor(df$MaritalStatus,levels = c("Single","Married","Divorced"))

ggplot(data=df,mapping=aes(x=Default,y=CreditScore)) +
  geom_violin()
```

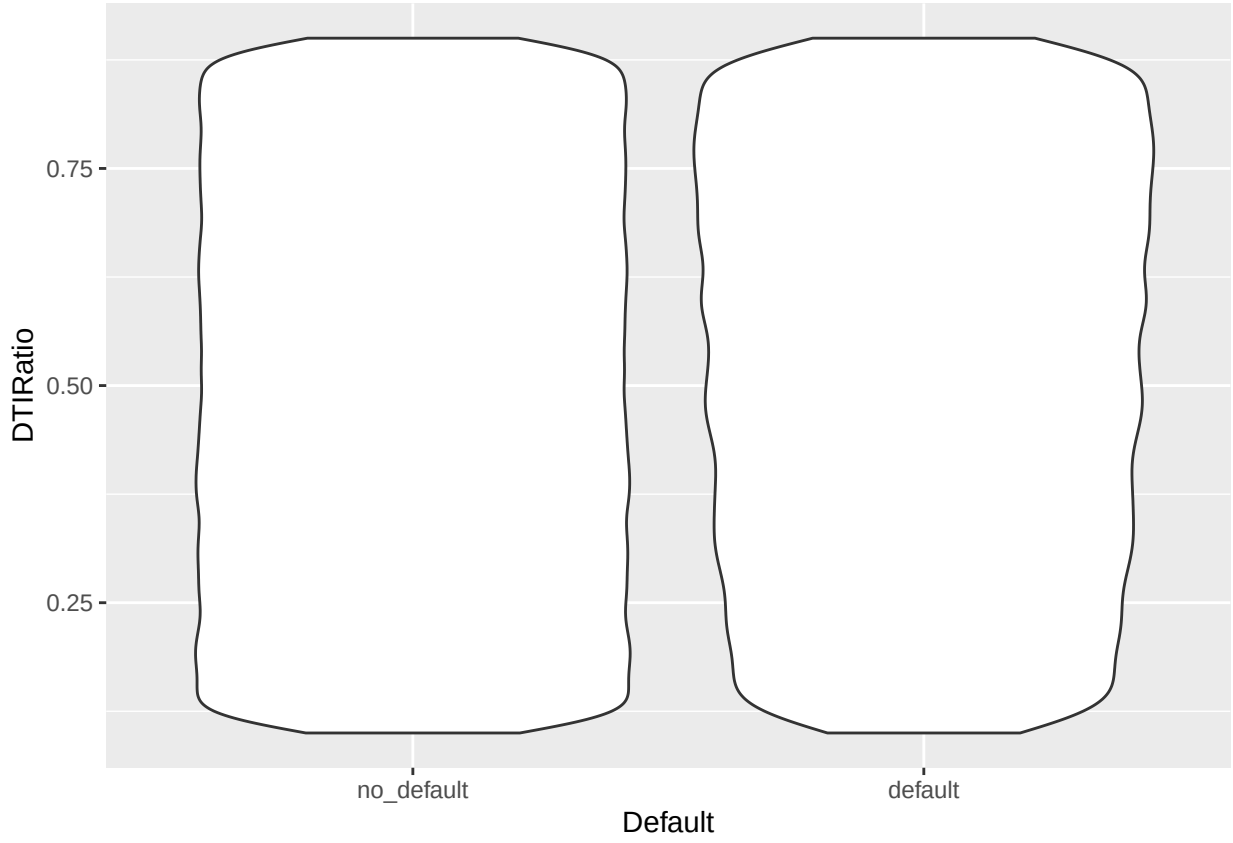



```r
t.test(df[df$Default=='default','CreditScore'],
       df[df$Default=='no_default','CreditScore'])
```

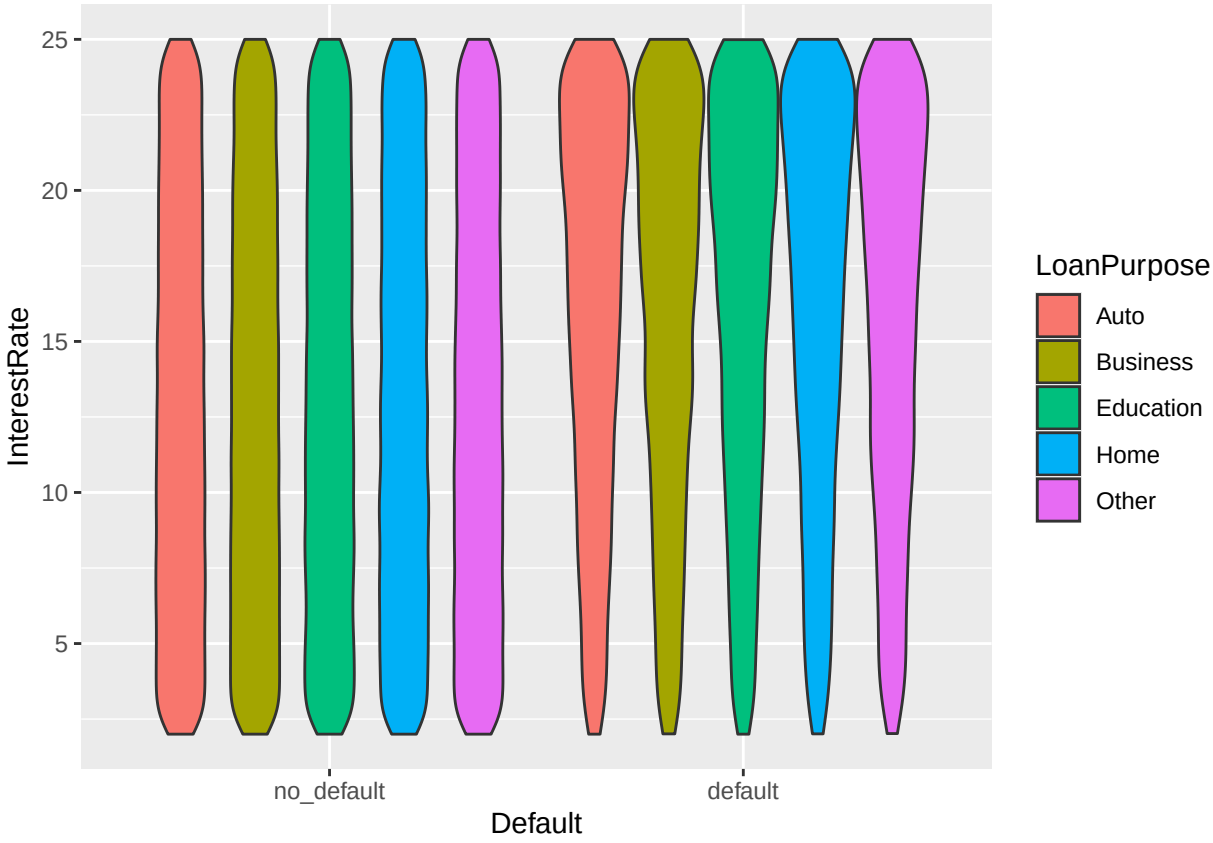
```
##
##  Welch Two Sample t-test
##
## data:  df[df$Default == "default", "CreditScore"] and df[df$Default == "no_default", "CreditScore"]
## t = -17.302, df = 37905, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.86580 -15.02646
## sample estimates:
```

```
## mean of x mean of y
##  559.2861  576.2323
```

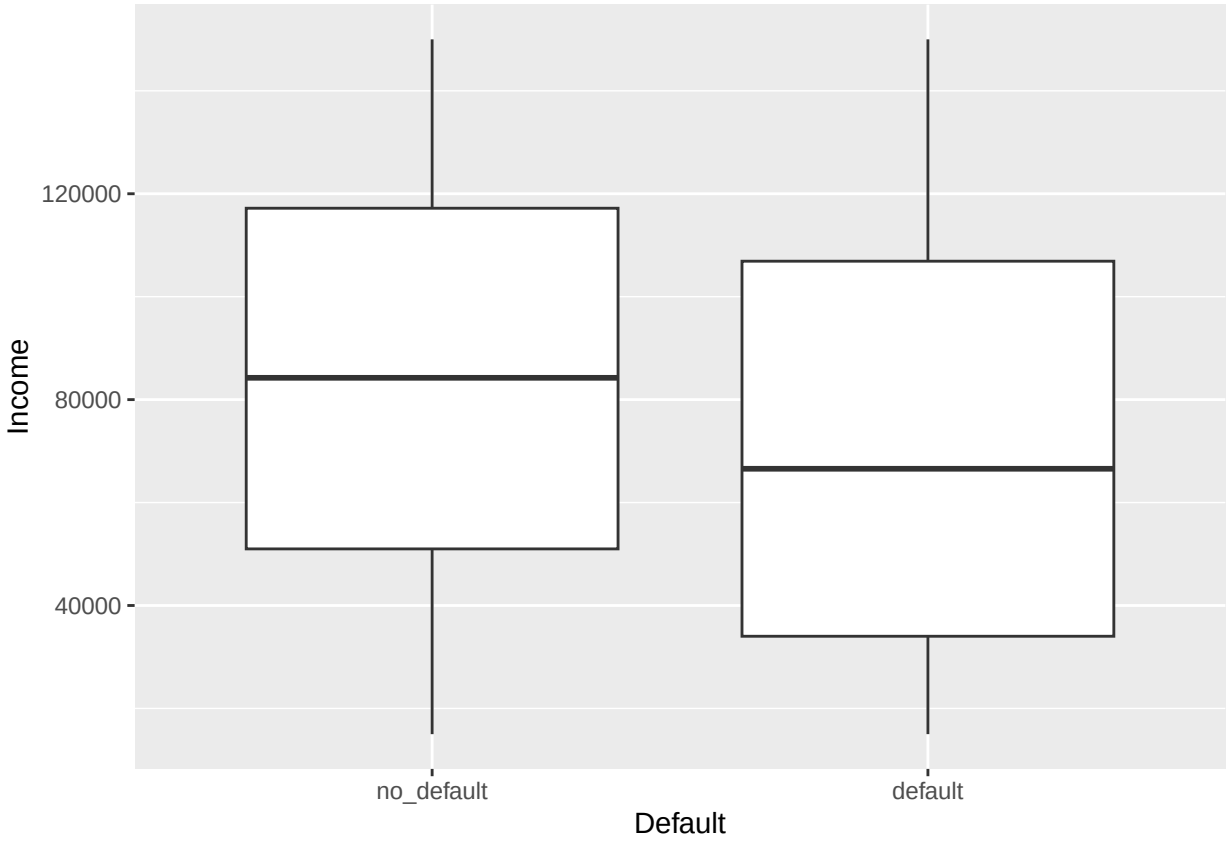
```
ggplot(data=df,mapping=aes(x=Default,y=DTIRatio)) +
  geom_violin()
```



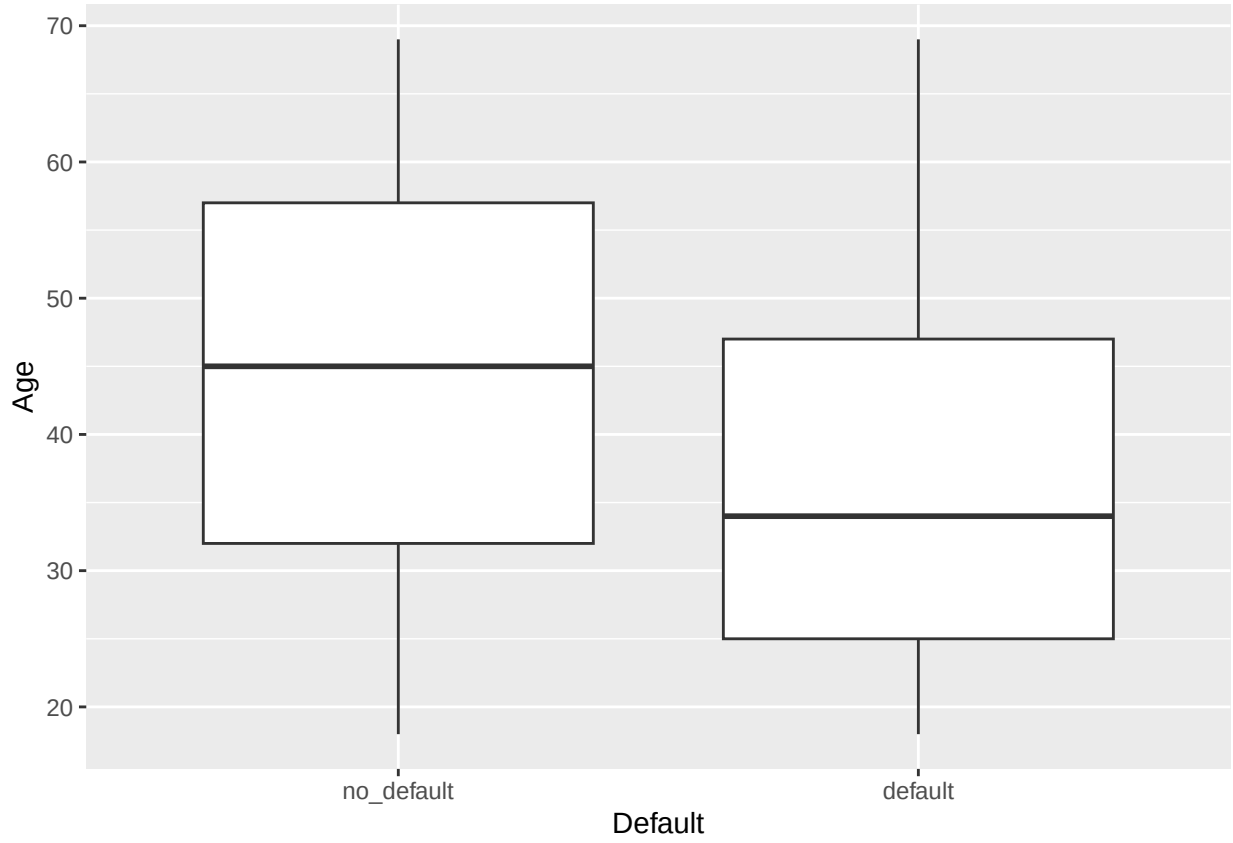

```
ggplot(data=df,mapping=aes(x=Default,y=InterestRate,fill=LoanPurpose)) +
  geom_violin()
```



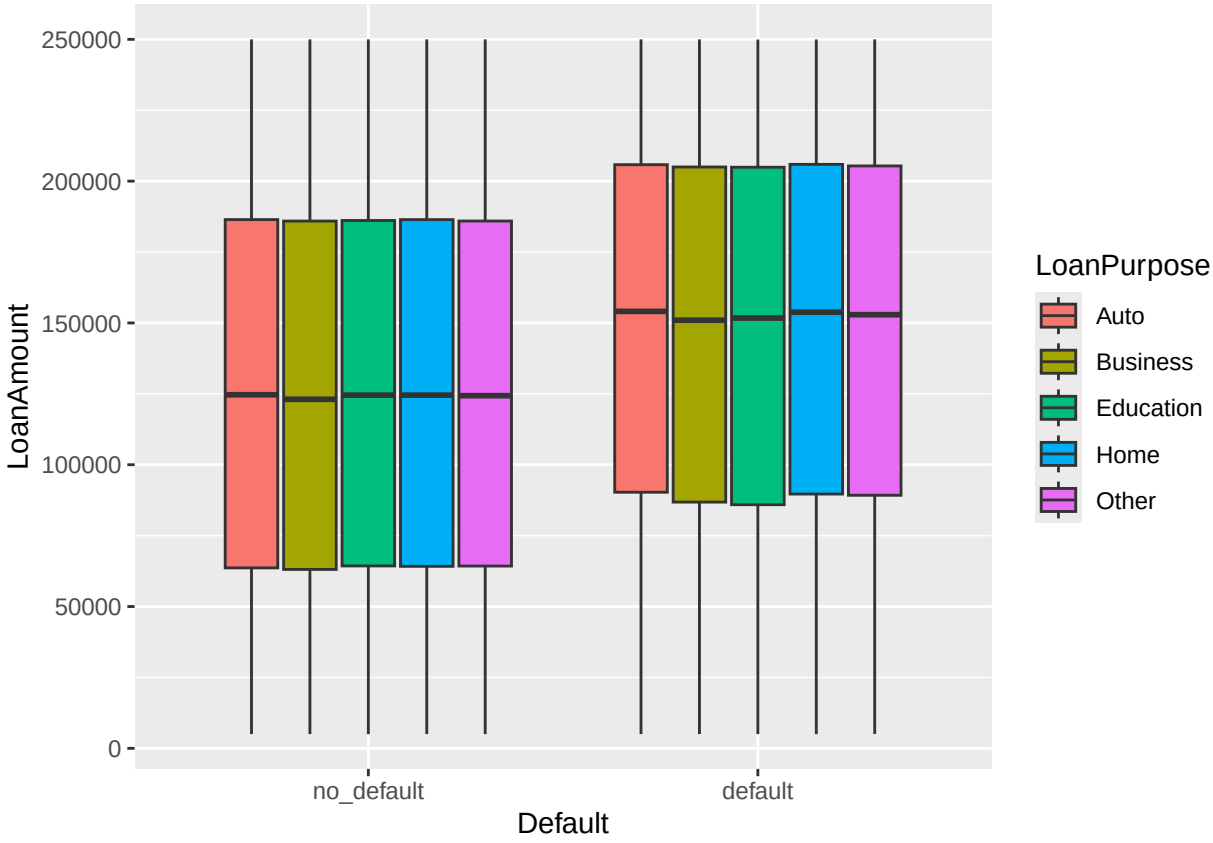

```
ggplot(data=df,mapping=aes(x=Default,y=Income)) +
  geom_boxplot()
```



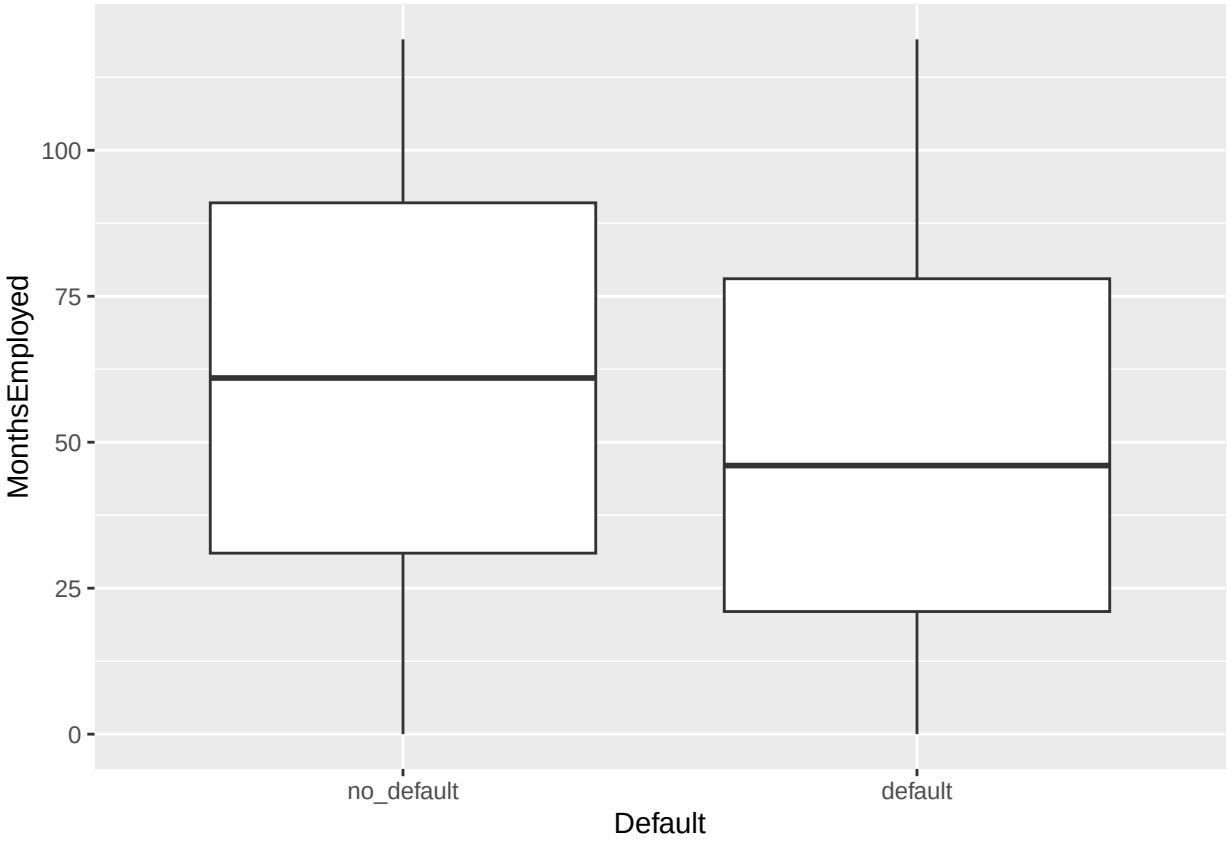

```
ggplot(data=df,mapping=aes(x=Default,y=Age)) +
  geom_boxplot()
```



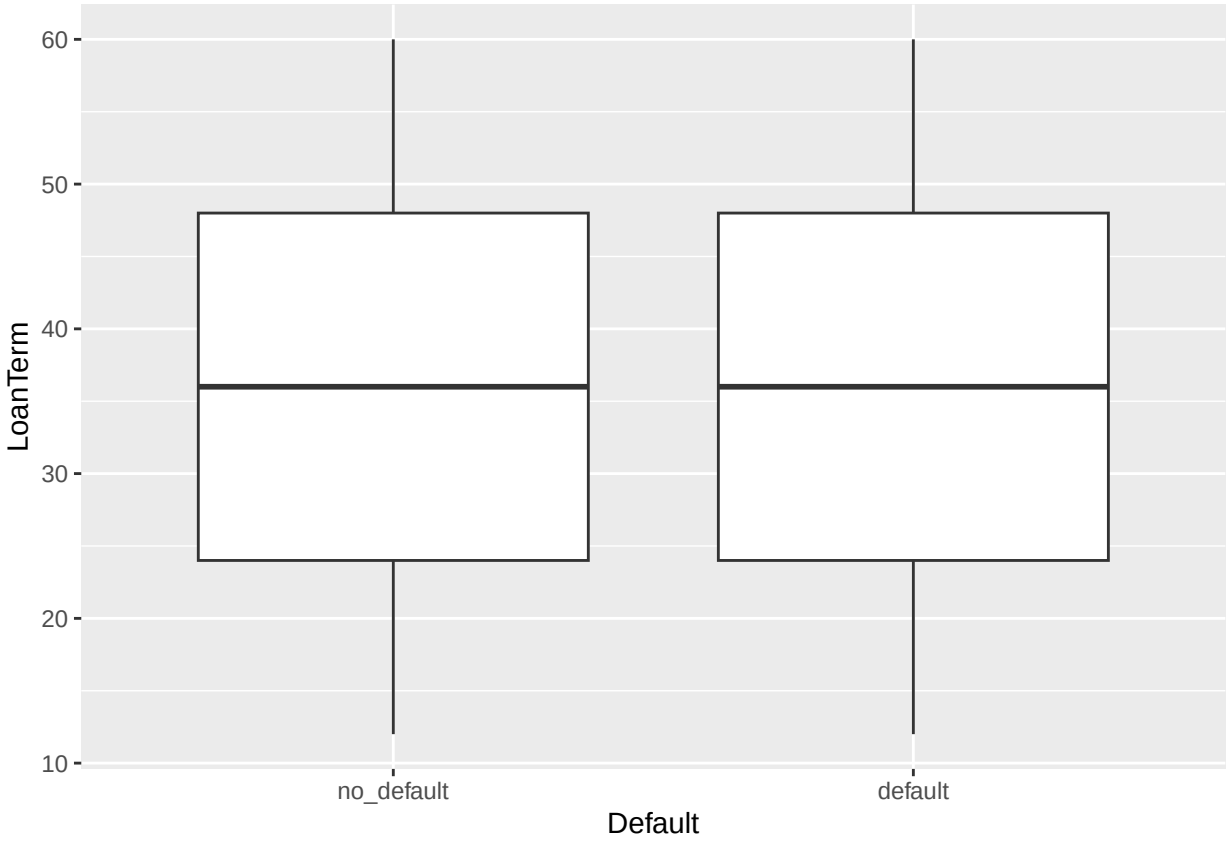

```
ggplot(data=df,mapping=aes(x=Default,y=LoanAmount,fill=LoanPurpose)) +
  geom_boxplot()
```




```
ggplot(data=df,mapping=aes(x=Default,y=MonthsEmployed)) +
  geom_boxplot()
```




```
ggplot(data=df,mapping=aes(x=Default,y=LoanTerm)) +
  geom_boxplot()
```




```
odds <- as.data.frame.matrix(table(df$Education,df$Default))
odds$odds <- odds$no_default / odds$default
print(odds)
```

```
##             no_default default     odds
## High School      55673    8230 6.764642
## Bachelor's       56577    7789 7.263705
## Master's         56633    6908 8.198176
## PhD              56811    6726 8.446476
```

```
odds <- as.data.frame.matrix(table(df$LoanPurpose,df$Default))
odds$odds <- odds$no_default / odds$default
print(odds)
```

```
##           no_default default     odds
## Auto           44803    6041 7.416487
## Business       44975    6323 7.112921
## Education      44967    6038 7.447334
## Home           46037    5249 8.770623
## Other          44912    6002 7.482839
```

```
odds <- as.data.frame.matrix(table(df$HasCoSigner,df$Default))
odds$odds <- odds$no_default / odds$default
print(odds)
```

```
##     no_default default     odds
## No      111223   16423 6.772392
## Yes     114471   13230 8.652381
```

```
odds <- as.data.frame.matrix(table(df$HasMortgage,df$Default))
odds$odds <- odds$no_default / odds$default
print(odds)
```

```
##     no_default default     odds
## No      111909   15761 7.100374
## Yes     113785   13892 8.190685
```

```
odds <- as.data.frame.matrix(table(df$EmploymentType,df$Default))
odds$odds <- odds$no_default / odds$default
print(odds)
```

```
##               no_default default     odds
## Full-time          57632    6024 9.567065
## Part-time          56484    7677 7.357562
## Self-employed      56404    7302 7.724459
## Unemployed         55174    8650 6.378497
```

```
odds <- as.data.frame.matrix(table(df$NumCreditLines,df$Default))
odds$odds <- odds$no_default / odds$default
print(odds)
```

```
##   no_default default     odds
## 1      56866    6688 8.502691
## 2      57038    7092 8.042583
## 3      56222    7612 7.385970
## 4      55568    8261 6.726546
```

```
odds <- as.data.frame.matrix(table(df$MaritalStatus,df$Default))
odds$odds <- odds$no_default / odds$default
print(odds)
```

```
##          no_default default     odds
## Single        74885   10127 7.394589
## Married       76433    8869 8.617995
## Divorced      74376   10657 6.979075
```

```
odds <- as.data.frame.matrix(table(df$HasDependents,df$Default))
odds$odds <- odds$no_default / odds$default
print(odds)
```

```
##     no_default default     odds
## No      111368   16237 6.858903
## Yes     114326   13416 8.521616
```