104_project3

Krish Methi, Krithik J, Brandon Chan

3/9/2024

```
rm(list=ls(all=TRUE))
# load libraries
library(AER)
```

```
## Warning: package 'AER' was built under R version 4.1.2
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.1.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.2
```

```
## Loading required package: lmtest
```

```
## Warning: package 'lmtest' was built under R version 4.1.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.1.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
library(plm)
```

```
## Warning: package 'plm' was built under R version 4.1.2
```

```
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 4.1.2
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(sandwich)
library(lmtest)
library(margins)
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.1.2
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.1.2
```

```
library(ggplot2)
library(car)
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

1. Panel Data Models

```
# Read in the panel data
Grunfeld <- read.csv("/Users/kmethi2/Downloads/Grunfeld.csv")
```

(a) Briefly discuss data and question answering

The data we are working with for our Panel data is the Grunfeld investment data. This data is a balanced panel of 11 large US manufacturing firms from the years 1935-1954. The variables include factor variables of firm and year, as well as Gross investment, market value of the firm, and the capital stock of the firm. The data has 220 observations of these 5 variables. We will be using this data to predict the market value of the firm, and we will make a pooled model, fixed and random effects model and see which model is the best fit to our panel data

(b) Give descriptive analysis of each variable

Investment

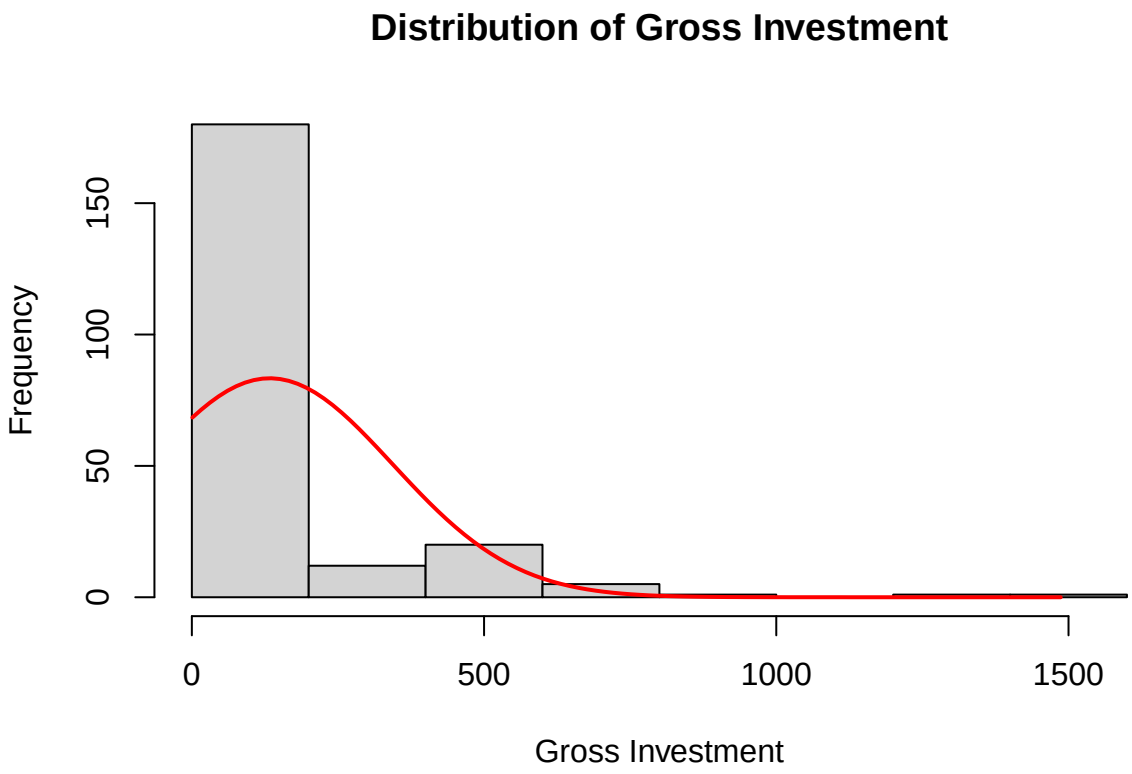
Gross investment is defined as the additions to plant and equipment plus maintenance and repairs in millions of dollars deflated by the implicit price deflator of producers' durable equipment (base 1947)

```
# statistical summary of the gross investment column
summary(Grunfeld$invest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.93   27.38   52.37  133.31   99.78 1486.70
```

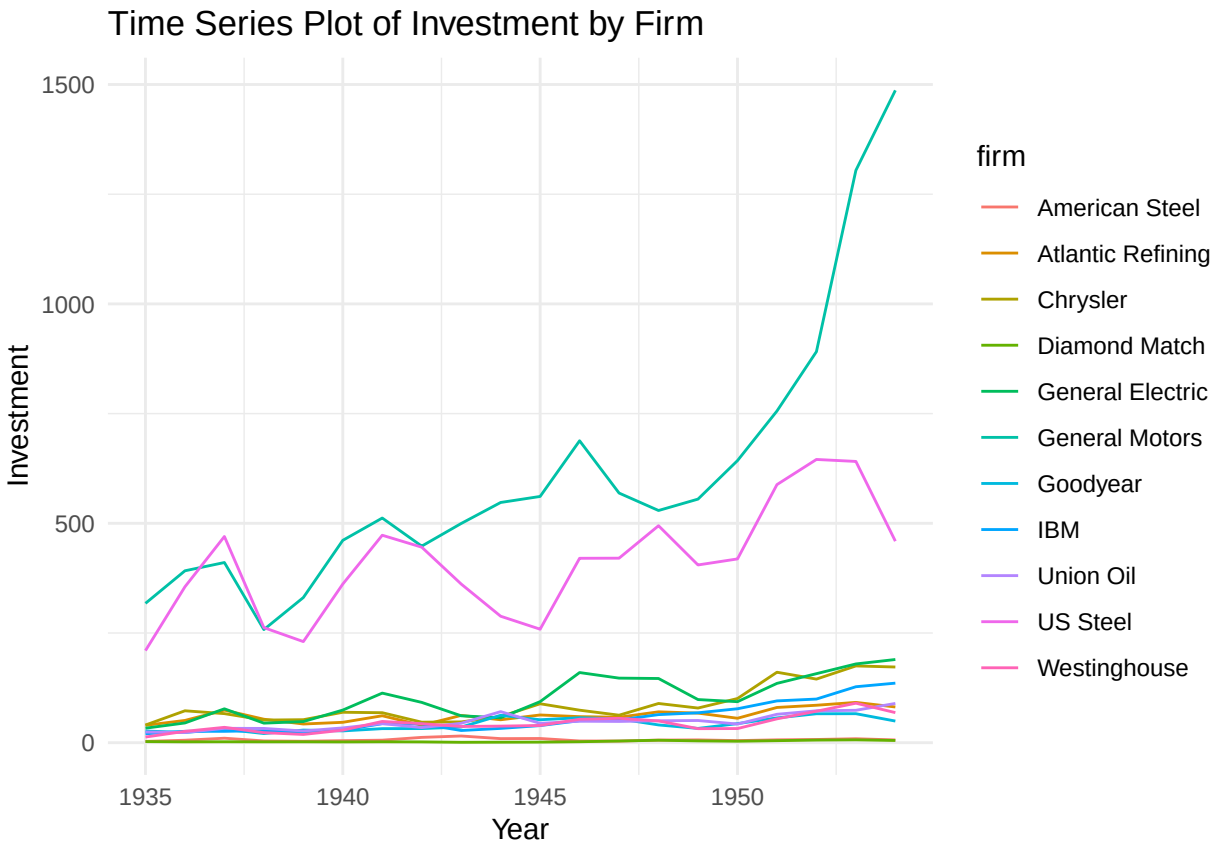
Here we have the statistical summary of the investment variable, we can see that there is quite a big spread from 0.93-1486.7, and the median and mean are 52.37 and 133.31, suggesting major right skewness.

```
# Fit the histogram and plot it with fitted normal curve
inv_hist <- hist(Grunfeld$invest, main = "Distribution of Gross Investment", xlab = "Gross Investment")
xfit <- seq(min(Grunfeld$invest), max(Grunfeld$invest), length = 100)
yfit <- dnorm(xfit, mean = mean(Grunfeld$invest), sd = sd(Grunfeld$invest))
yfit <- yfit * diff(inv_hist$mids[1:2]) * length(Grunfeld$invest)
lines(xfit, yfit, col = "red", lwd = 2)
```



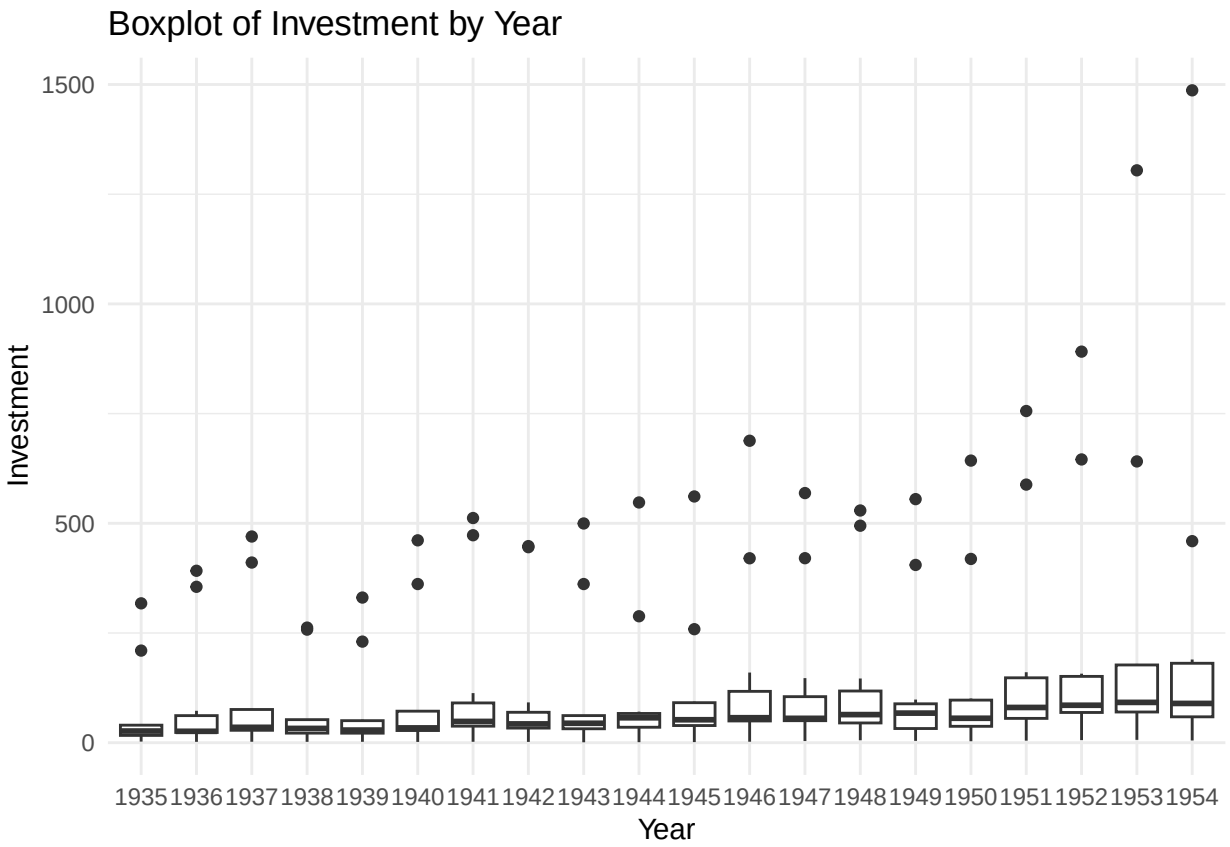
From the histogram, we can see that the investment data is heavily right skewed, and clearly does not overlay well with the normal curve.

```
# Fit a basic time series plot of investment by firm over time
ggplot(Grunfeld, aes(x = year, y = invest, color = firm)) +
  geom_line() +
  theme_minimal() +
  labs(title = "Time Series Plot of Investment by Firm", y = "Investment", x = "Year")
```



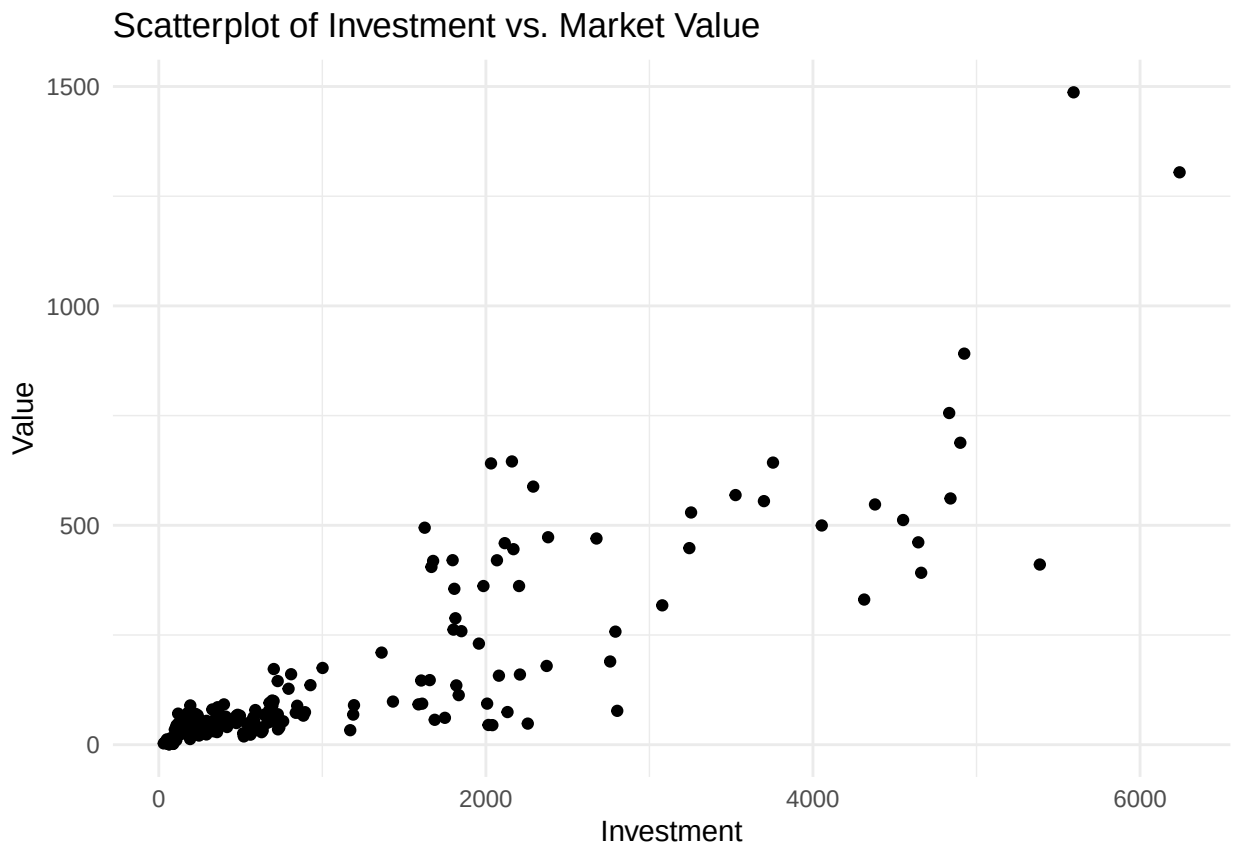
Here we have a time series plot of investment by firm. We can see that most firms are on the bottom in terms of investment and don't have much of a trend, however there are a couple of firms with different time series curves with higher values and trends. There also seems to be individual heterogeneity over time as well.

```
# Fit a boxplot year over year of investment
ggplot(Grunfeld, aes(x = as.factor(year), y = invest)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of Investment by Year",
       x = "Year",
       y = "Investment")
```



Here we can see boxplots of investment for every year in the data. We see that there are outliers on the higher side all across the data, with higher and higher outliers as the years go on.

```
# Plot a Basic scatterplot of investment vs market value
ggplot(Grunfeld, aes(x = value, y = invest)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Scatterplot of Investment vs. Market Value", x = "Investment", y = "Value")
```



Here we have a scatterplot of investment vs market value, our variable of interest. We see a pretty clear trend that as investment goes up, the market value of the firm also increases.

Market Value

The value column is defined as the Market value of the firm, defined as the price of common shares at December 31 (or, for WH, IBM and CH, the average price of December 31 and January 31 of the following year) times the number of common shares outstanding plus price of preferred shares at December 31 (or average price of December 31 and January 31 of the following year) times number of preferred shares plus total book value of debt at December 31 in millions of dollars deflated by the implicit GNP price deflator (base 1947).

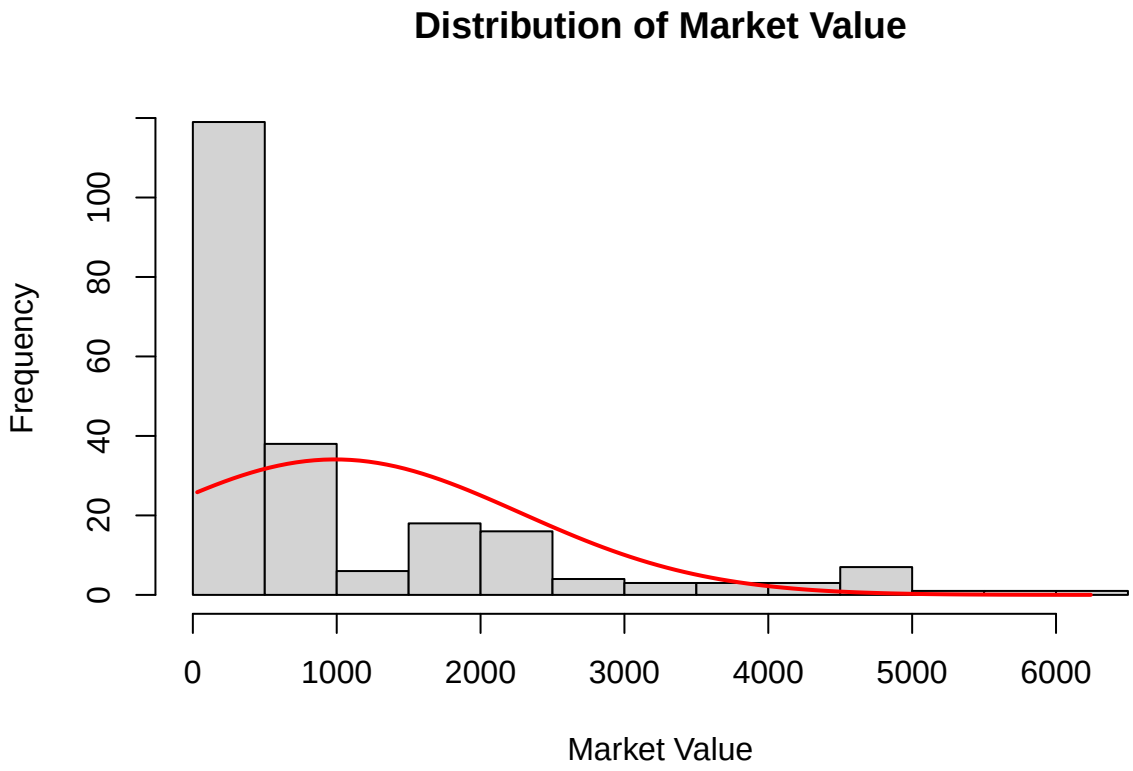
```
# statistical summary of the Market Value column
summary(Grunfeld$value)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.28  160.32  404.65  988.58 1605.92 6241.70
```

Here we have the statistical summary of the value variable, we can see that there is quite a big spread from 30.28-6241.7, and the median and mean are 404.65 and 988.58, suggesting major right skewness.

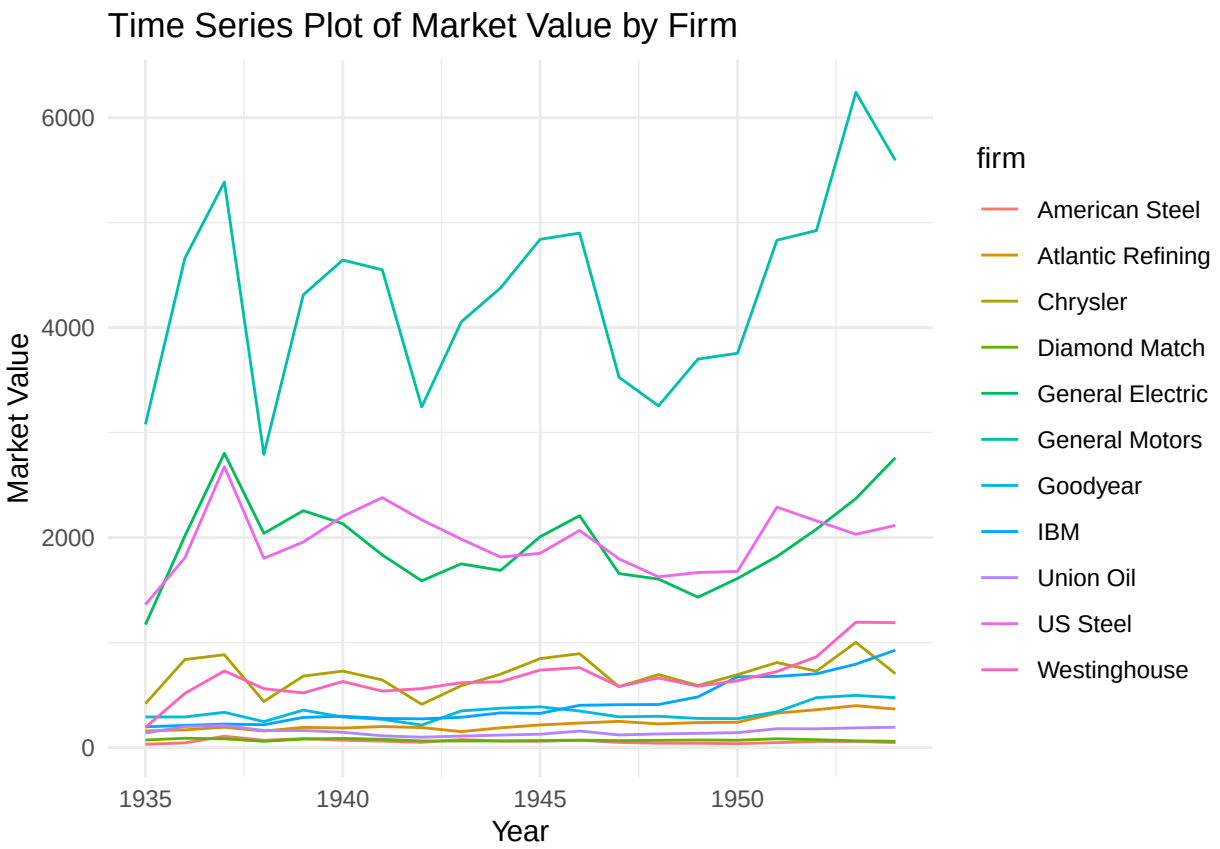
```
# Fit the histogram and plot it with fitted normal curve
value_hist <- hist(Grunfeld$value, main = "Distribution of Market Value", xlab = "Market Value")
xfit <- seq(min(Grunfeld$value), max(Grunfeld$value), length = 100)
yfit <- dnorm(xfit, mean = mean(Grunfeld$value), sd = sd(Grunfeld$value))
```

```
yfit <- yfit * diff(value_hist$mids[1:2]) * length(Grunfeld$value)
lines(xfit, yfit, col = "red", lwd = 2)
```



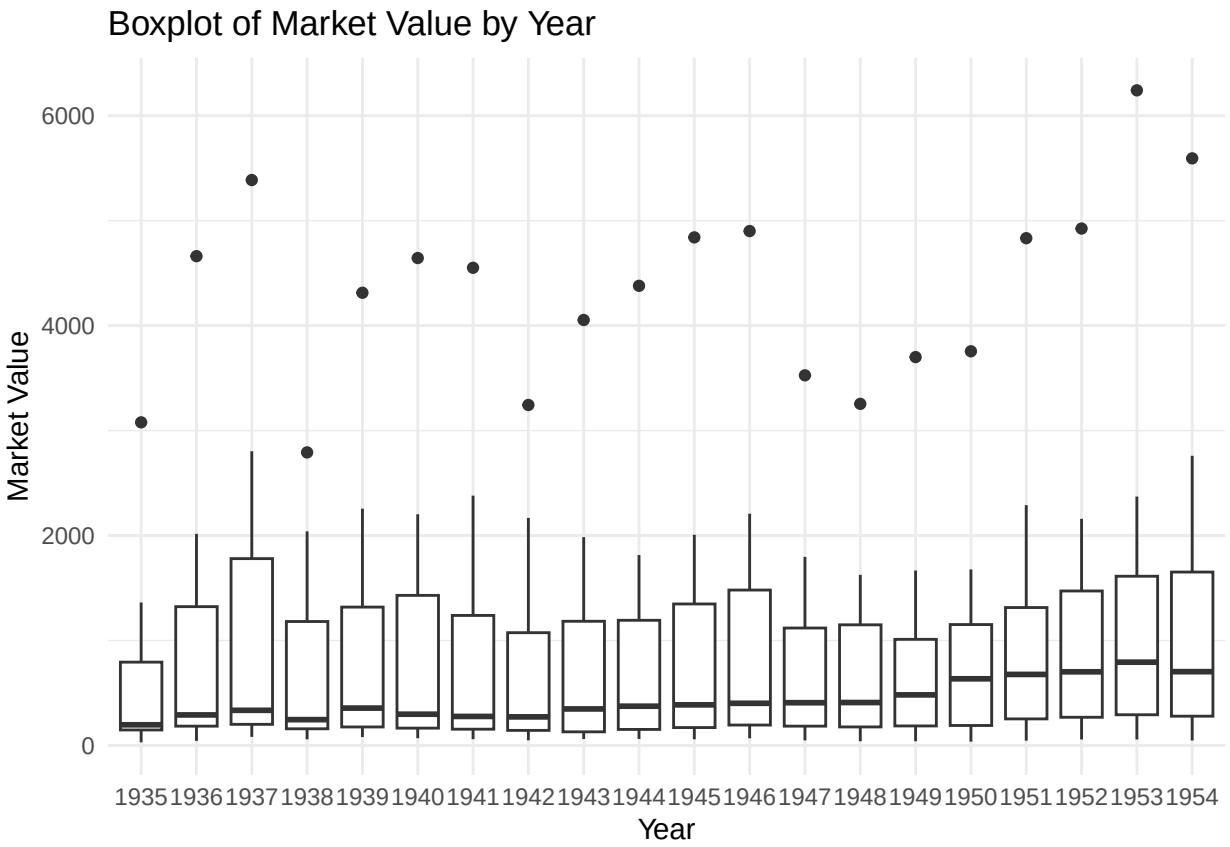
From the histogram, similar to the investment histogram, we can see that the market value data is also heavily right skewed, and clearly does not overlay well with the normal curve.

```
# Fit a basic time series plot of Market Value by firm over time
ggplot(Grunfeld, aes(x = year, y = value, color = firm)) +
  geom_line() +
  theme_minimal() +
  labs(title = "Time Series Plot of Market Value by Firm", y = "Market Value", x = "Year")
```



Here we have the time series plot by firm of market value over time. We can see that there are a cluster of firms that have low relative market value and no big trend, but again there are certain firms with higher market values that have more of a positive trend, potentially the same firms as in investment. There also seems to be individual heterogeneity present as well.

```
# Fit a boxplot year over year of Market Value
ggplot(Grunfeld, aes(x = as.factor(year), y = value)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of Market Value by Year",
       x = "Year",
       y = "Market Value")
```



Here we can see boxplots of Market value for every year in the data. We see that there are outliers on the higher side all across the data, however they seem to be more stationary than the investment column in terms of of their scale except for the most recent couple yeas. The variation seems to increase year over year from the boxplot for market value.

Capital

The capital column is defined as the Stock of plant and equipment, defined as the accumulated sum of net additions to plant and equipment deflated by the implicit price deflator for producers' durable equipment (base 1947) minus depreciation allowance deflated by depreciation expense deflator (10 years moving average of wholesale price index of metals and metal products, base 1947).

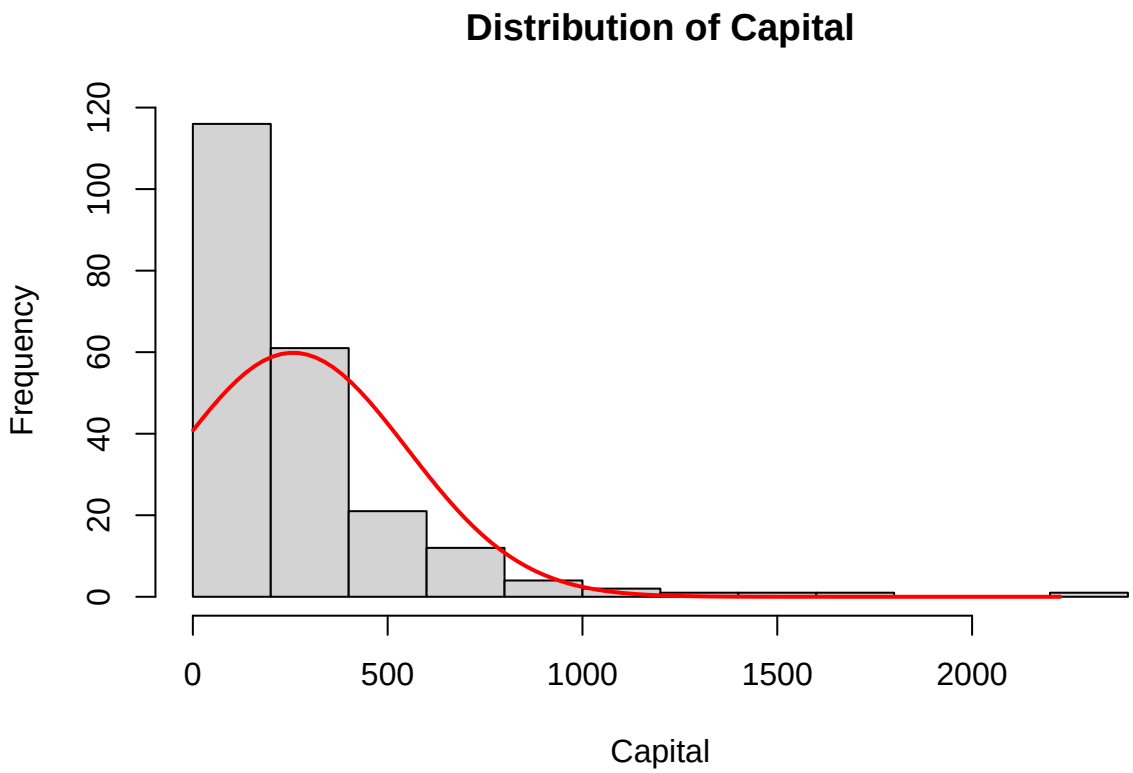
```
# statistical summary of the capital column
summary(Grunfeld$capital)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.8    67.1   180.1   257.1   344.5  2226.3
```

Here we have the statistical summary of the capital variable, with a big spread from 0.8-2226.3, and the median and mean are 180.1 and 257.1, suggesting right skewness again however at a relatively lower scale than the previous variables

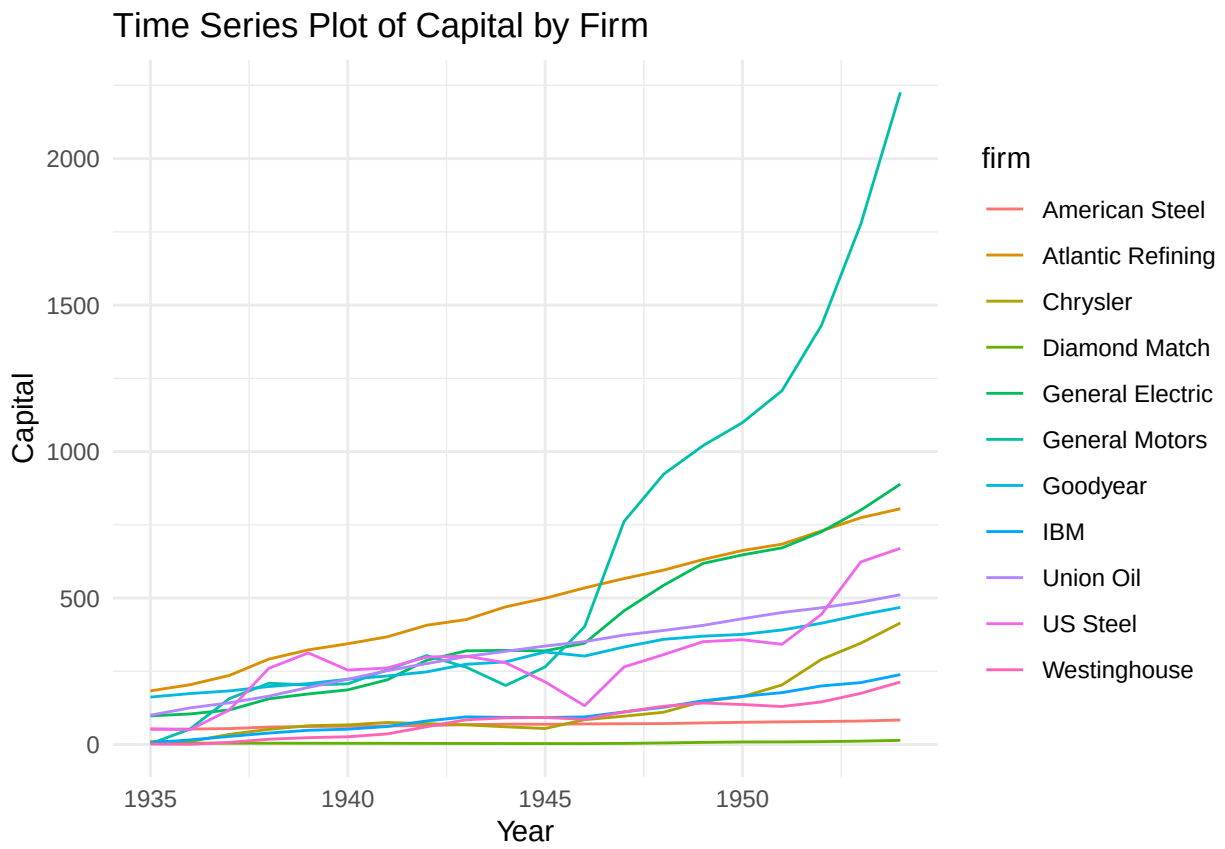
```
# Fit the histogram and plot it with fitted normal curve
cap_hist <- hist(Grunfeld$capital, main = "Distribution of Capital", xlab = "Capital")
xfit <- seq(min(Grunfeld$capital), max(Grunfeld$capital), length = 100)
```

```
yfit <- dnorm(xfit, mean = mean(Grunfeld$capital), sd = sd(Grunfeld$capital))
yfit <- yfit * diff(cap_hist$mids[1:2]) * length(Grunfeld$capital)
lines(xfit, yfit, col = "red", lwd = 2)
```



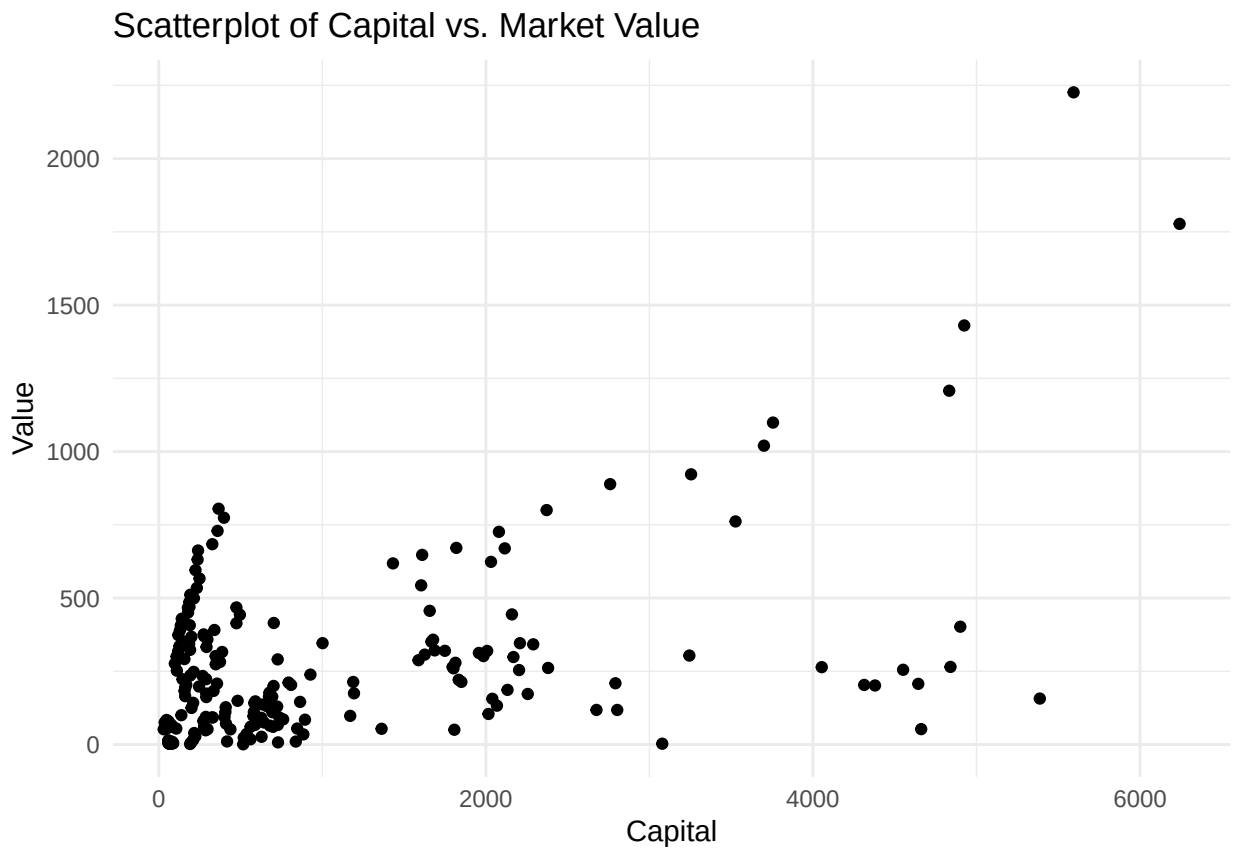
From the histogram, similar to the other histograms, we can see that the capital data is also heavily right skewed, and clearly does not overlay well with the normal curve.

```
# Fit a basic time series plot of capital by firm over time
ggplot(Grunfeld, aes(x = year, y = capital, color = firm)) +
  geom_line() +
  theme_minimal() +
  labs(title = "Time Series Plot of Capital by Firm", y = "Capital", x = "Year")
```



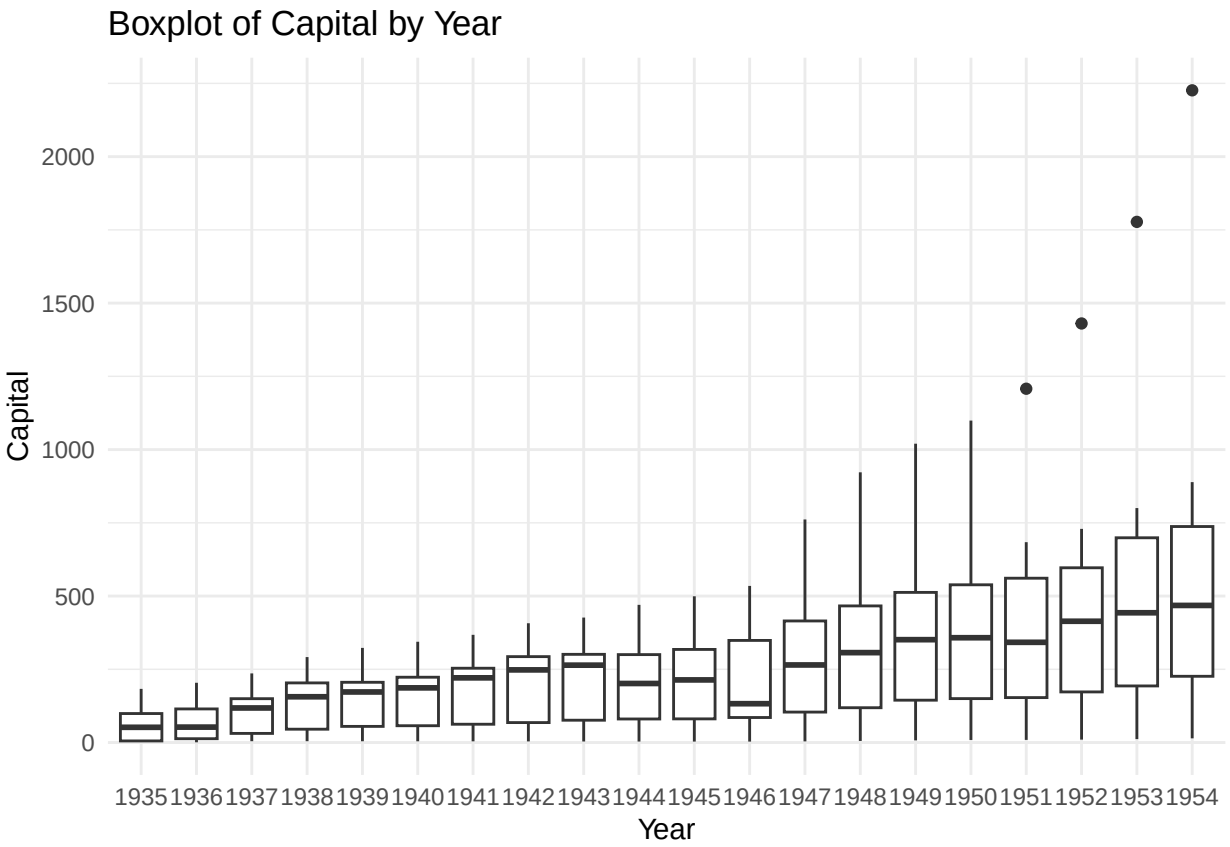
Here we have the time series plot by firm of capital over time. Here we see a different plot than the previous two. Here, it looks pretty similar firm by firm over time, except for one firm that has a massive positive exponential like increase towards the end of the timeframe. Other than that, the data looks pretty similar between firms, with slight positive trends over time for the most part. There does also seem to be individual heterogeneity present although not as much as the other two variables.

```
# Plot a Basic scatterplot of capital vs market value
ggplot(Grunfeld, aes(x = value, y = capital)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Scatterplot of Capital vs. Market Value", x = "Capital", y = "Value")
```



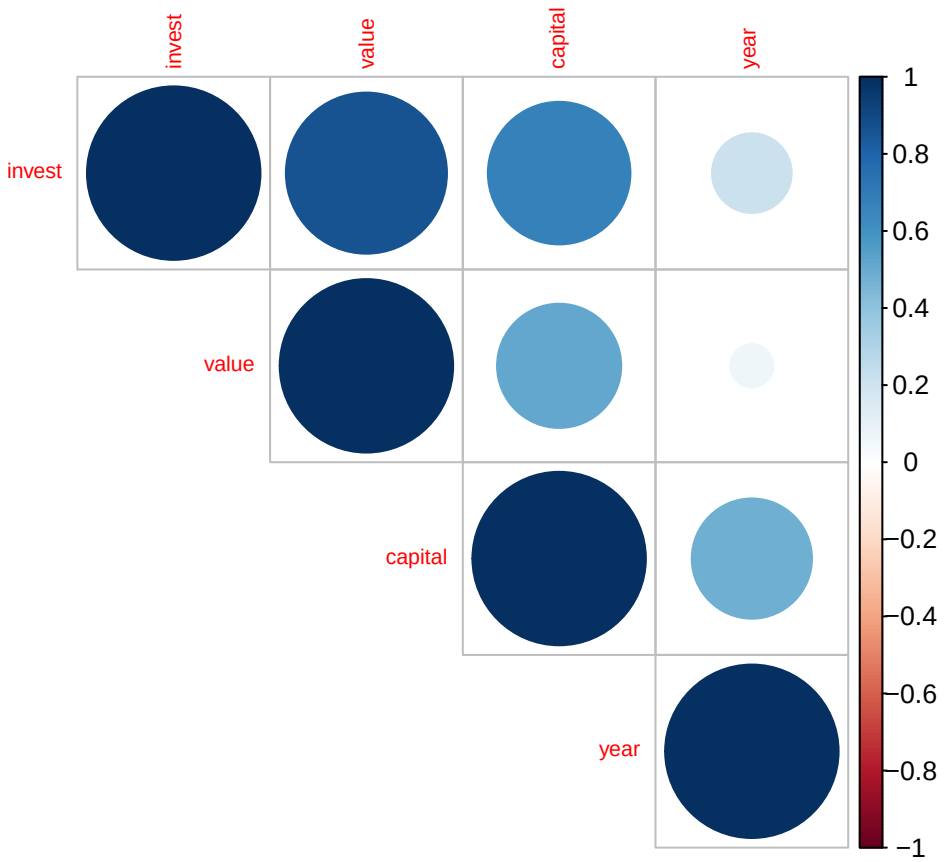
Here we have a scatterplot of capital vs market value, our variable of interest. Here the data also is a bit different, we don't see as much of a big positive association between capital and market value, but there definitely is a positive association.

```
# Fit a boxplot year over year for capital
ggplot(Grunfeld, aes(x = as.factor(year), y = capital)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of Capital by Year",
       x = "Year",
       y = "Capital")
```



Here we can see boxplots of capital for every year in the data. We see that there are outliers on the higher side only towards the end of the data. For the most part, the boxplots year over year seem to be pretty stationary, except for towards the end of the data. Similar to that, the variation stays relatively constant until we get to around 1946-47, coincidentally right around World War 2, which may have influenced that development. in terms of the outliers, we also see there are not nearly as many outliers as the other two variables, another component showing the more stationary nature over time of capital.

```
# Look at the correlations between variables in our dataset
panel_correlation_matrix <- cor(Grunfeld[sapply(Grunfeld, is.numeric)])
corrplot(panel_correlation_matrix, method = "circle", type = "upper", tl.cex = 0.7)
```



Because the firm column was a categorical variable with the names of the firm, it was excluded from the correlation plot. Based on this plot, we can see the largest correlation is Market Value to investment, and we can see the smallest correlation is year to market value. Economically, this makes sense with intuition as your firm is growing, your investment will also most likely grow. Overall, we see pretty high correlations across the board except for a couple correlations involving year, which again is what you expect as investment, capital and market value are all very much related economically.

Pooled model

```
# Convert data to panel structure and create pooled model
panel_data <- pdata.frame(Grunfeld, index=c("firm", "year"))
pooled_model <- plm(value~invest+capital, model="pooling", data=panel_data)
summary(pooled_model)
```

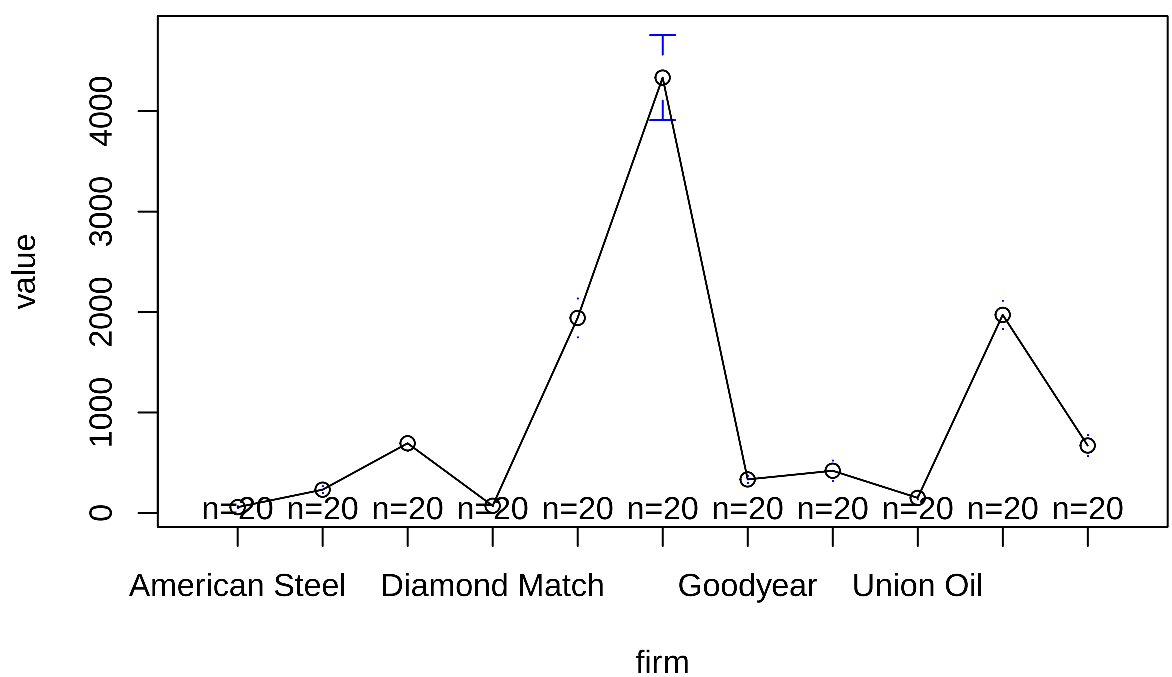
```
## Pooling Model
##
## Call:
## plm(formula = value ~ invest + capital, data = panel_data, model = "pooling")
##
## Balanced Panel: n = 11, T = 20, N = 220
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -2135.686  -302.521  -183.627    88.597  2731.793
##
```

```
## Coefficients:
##              Estimate Std. Error t-value  Pr(>|t|)
## (Intercept) 360.41243   57.74234  6.2417 2.243e-09 ***
## invest        5.80585    0.27975 20.7534 < 2.2e-16 ***
## capital      -0.56717    0.20091 -2.8230    0.0052 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    362910000
## Residual Sum of Squares: 89656000
## R-Squared:      0.75296
## Adj. R-Squared: 0.75068
## F-statistic: 330.693 on 2 and 217 DF, p-value: < 2.22e-16
```

```r
# Plotmeans and BP test
plotmeans(value ~ firm, data = panel_data)
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

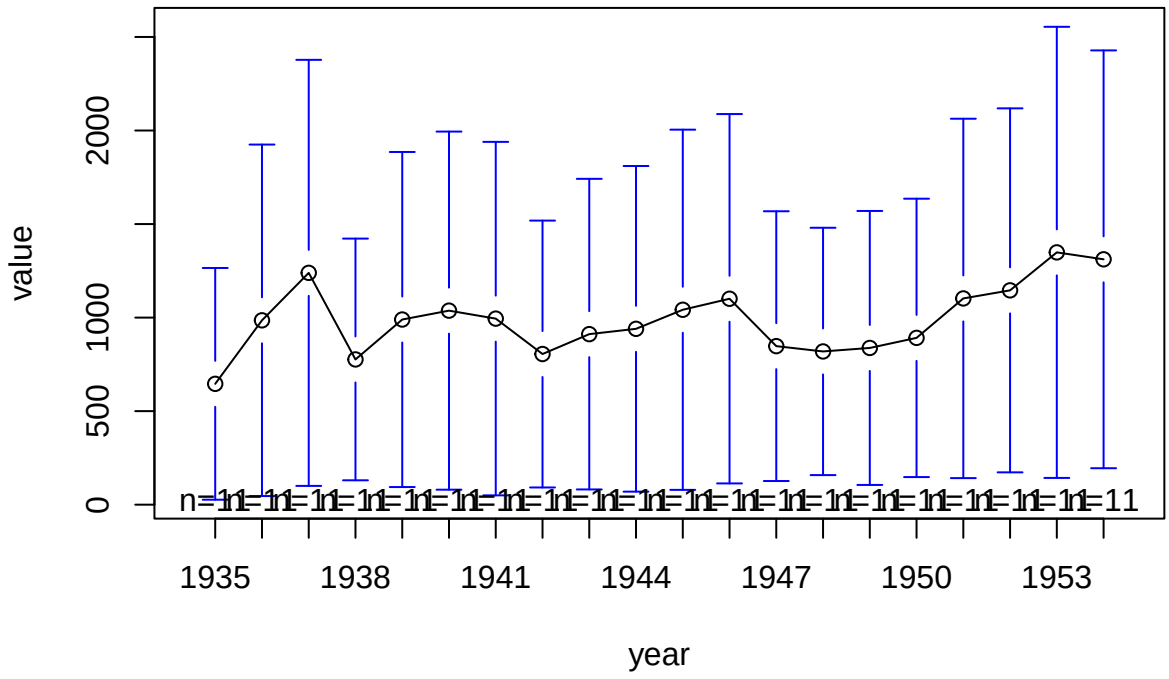
```
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```



```
bptest(value ~ firm, data = panel_data)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  value ~ firm
## BP = 79.382, df = 10, p-value = 6.636e-13
```

```
plotmeans(value ~ year, data = panel_data)
```




```
bptest(value ~ year, data = panel_data)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  value ~ year
## BP = 6.2546, df = 19, p-value = 0.9973
```

```
#BP test on pooled model and plot of residuals vs fitted values
bptest_result <- bptest(pooled_model)
bptest_result
```

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  pooled_model
## BP = 43.64, df = 2, p-value = 3.339e-10
```

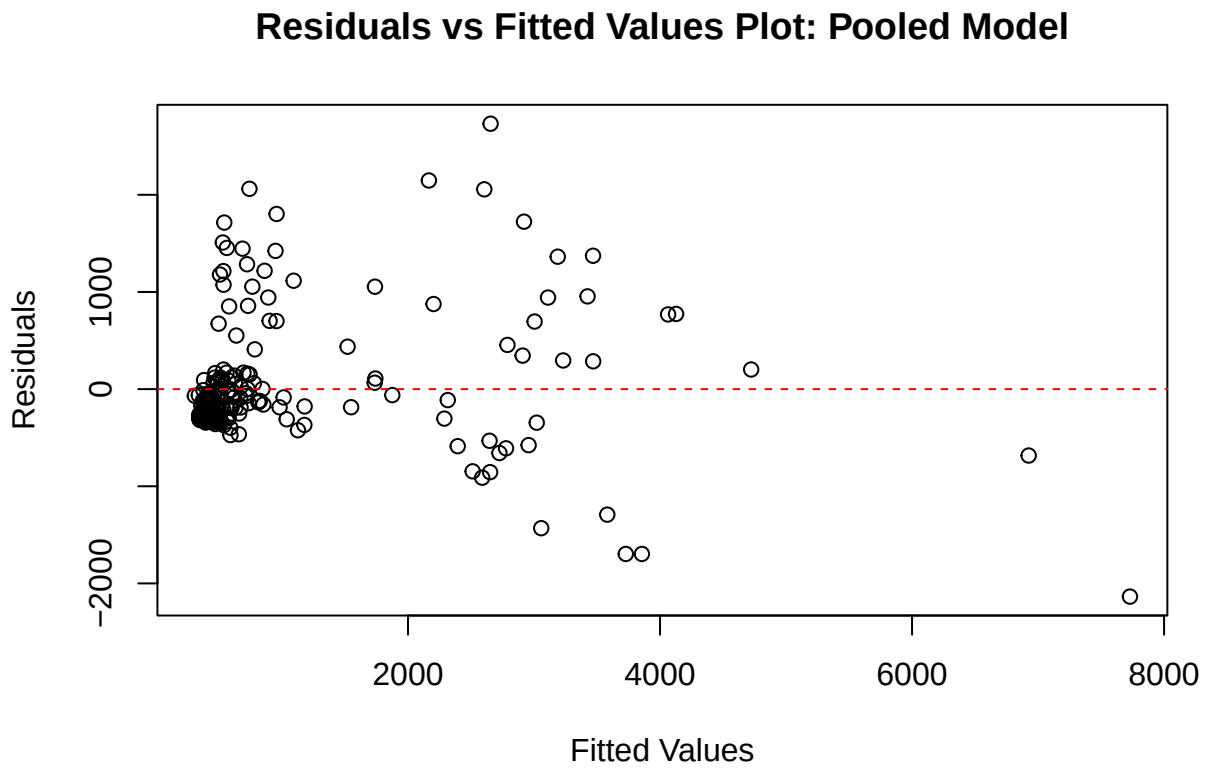
```
residuals_pooled <- residuals(pooled_model)
fitted_values_pooled <- fitted(pooled_model)
head(fitted_values_pooled)
```

```
## [1] 347.9710 363.1700 388.9136 350.0305 344.7517 352.2816
```

```
head(residuals_pooled)
```

```
## American Steel-1935 American Steel-1936 American Steel-1937 American Steel-1938
##           -317.6870           -319.2610           -281.8936           -281.7245
## American Steel-1939 American Steel-1940
##           -260.5877           -283.1246
```

```
df_new = as.data.frame(cbind(residuals_pooled, fitted_values_pooled))
plot(df_new$fitted_values_pooled, df_new$residuals_pooled, ylab = "Residuals", xlab = "Fitted Values", m
abline(h = 0, col = "red", lty = 2)
```



Based on the mean plots and the BP test, in terms of heterogeneity, we can see that it's pretty evident that there is evidence for individual heterogeneity, but not that for heterogeneity over time. The P value of the BP test was extremely small for the individual, however extremely large for the year.

Looking at the pooled model, the plot of the residuals vs fitted values appears to not have any big pattern, however after looking at the P value, it is extremely low so we reject the null hypothesis and conclude there is heteroskedasticity in the pooled model. Therefore we should correct for that using cluster robust standard errors.

So overall, we see that there seems to be individual heterogeneity but not time heterogeneity, and the pooled model has evidence of heteroskedasticity.

Fixed Effects Model

```
#Fixed effects model with only firm effects
fixed_firm <- plm(value ~ invest + capital, data = panel_data, model = "within", effect = "individual")
#Fixed effects model with both firm and time effects
fixed_both <- plm(value ~ invest + capital, data = panel_data, model = "within", effect = "twoways")
#Fixed effects model with only time effects
fixed_time <- plm(value ~ invest + capital, data = panel_data, model = "within", effect = "time")
##Test for fixed effects or pooled model
# Case 1:both time and individual effects jointly
pFtest(fixed_both, pooled_model)
```

```
##
##  F test for twoways effects
##
## data:  value ~ invest + capital
## F = 50.526, df1 = 29, df2 = 188, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
# Case 2: only time effects
pFtest(fixed_time,pooled_model)
```

```
##
##  F test for time effects
##
## data:  value ~ invest + capital
## F = 0.62386, df1 = 19, df2 = 198, p-value = 0.8856
## alternative hypothesis: significant effects
```

```
# Case 3: only individual effects
pFtest(fixed_firm,pooled_model)
```

```
##
##  F test for individual effects
##
## data:  value ~ invest + capital
## F = 115.94, df1 = 10, df2 = 207, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
# compare adjusted r^2 between selected models
summary(fixed_firm)$r.squared
```

```
##       rsq    adjrsq
## 0.4116245 0.3775158
```

```
summary(fixed_both)$r.squared
```

```
##       rsq    adjrsq
## 0.3657105 0.2611202
```

After calculating the fixed effects model and comparing that to the pooled model by use of the pF test, we can see that the test rejects the null for both the twoway fixed effects model, and the individual effects model, so in those cases, the fixed effects model is preferred over the pooled model. Since we want to get one fixed effect model, we compared the adjusted R^2 between the fixed twoway effect and the fixed firm effect. We can see that the fixed effects model with only firm had a better adjusted R^2, so that is the preferred model we would go with as of now. Here are our more formal interpretations of the pF tests:

F test for twoways effects: We reject the null since the p-value is tiny (< 2.2e-16), indicating that both individual and time effects are significant. Therefore, there is variation in the dependent variable that can be attributed to individual (country) and time effects. This tells us that fixed effects are needed and that the fixed_both model is suggested since pooled estimator is insufficient.

F test for time effects: Fail to reject the null as the p-value (0.8856) is greater than 0.05. So, fixed_time model which is only time effects is not recommended as time effects are not found to significantly contribute to the model.

F test for firm(individual) effects: We reject the null since the p-value(< 2.2e-16) is less than 0.05, indicating that individual (country) effects are significant. The fixed_firm model is preferred over pooled model.

From our F tests we know firm effects are significant but also both firm and time together are significant. If we compare their adjusted R^2 we conclude that fixed_firm has a higher value which could suggest that the model with only firm effects provides a slightly better fit to the data compared to the model with both firm and time effects.

Random Effects Model

```
# Create the random effects model, and run plm and ph tests to determine validity of model and model pr
random_model <- plm(value ~ invest + capital, data = panel_data, model = "random")
# plm test to determine necessity of the random effects model
random_effects_test <- plmtest(random_model, effect = "individual")
random_effects_test
```

```
##
##  Lagrange Multiplier Test - (Honda)
##
## data:  value ~ invest + capital
## normal = 28.673, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
# ph test to decide between the fixed model and the random effects model
hausman_test <- phtest(fixed_firm, random_model)
hausman_test
```

```
##
##  Hausman Test
##
## data:  value ~ invest + capital
## chisq = 144.89, df = 2, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

For the LM test we get a p-value of around 0. Rejecting the null implies a non-zero variance. By rejecting the null, this means that there's statistical evidence that individual heterogeneity exists and thus validating the use of our random effects model and not the pooled model.

The Hausman test is used to decide which is better: the Fixed Effects Model or the Random Effects Model. For the Hausman test, the null here implies that individual random effects are exogenous and that REM is better. With a p-value of ~0, we reject the null, meaning that we have evidence to suggest that the coefficients estimated by the FEM and the REM are statistically different, telling us to use our fixed firm effects model.

So overall, our conclusion from our testing and model fitting is that our fixed firm effects model is our preferred model. We came to this result by fitting a pooled, fixed effect, and random effect model and then running tests between the models to figure out preference. First, we concluded that a pooled model was not preferred over the other two models by the BP test, and then also confirmed by the pF and plm tests. Then, we decided which fixed model was the best, and the pF test and adjusted R^2 pointed us to the fixed firm effects model, and then lastly by running the Hausman test, we determined the fixed effects model was our overall preferred model.

Looking back, this conclusion makes sense as when looking at descriptions of the variables, we saw how for Investment, Capital and Market Value, there really was not much dynamics over time, but much more dynamics and heterogeneity by firm so it makes sense that the preferred model is the one that accounts for that, with the fixed effects firm model. However, an important thing to note is that both the Random and Fixed effects model were preferred at a point, so utilizing a Hausman Taylor estimator may also be something to think about using with this data.

Part 2 - Qualitative Dependent Variable Models

Briefly discuss your data and economic/finance/business question you are trying to answer with your model

Using all the predictors, we want to predict if the individual is a foreigner (this takes place in Switzerland, so are they Swiss or not?). The variables that we will be using to determine this are participation (did this person participate in the workforce), income, age in decades, years of education, number of kids under 7 (youngkids), and number of kids over 7 (oldkids).

Provide a descriptive analysis of your variables. This should include histograms and fitted distributions, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments.

The variable "foreign" is whether the person is a foreigner (if yes, it means the person is Swiss). The variable "participation" is whether the person participated in the labor force (yes or no). The variable "income" is the logarithm of nonlabor income. The variable "age" is age in decades. The variable "education" is the years of formal education. The variable "youngkids" is the number of young children (under 7 years old). The variable "oldkids" is the number of older children (over 7 years old).

```
# read in libraries and read in our dataset
library(ggplot2)
library(corrplot)
data("SwissLabor")
```

Our correlation plot is at the bottom below all the variables' graphs.

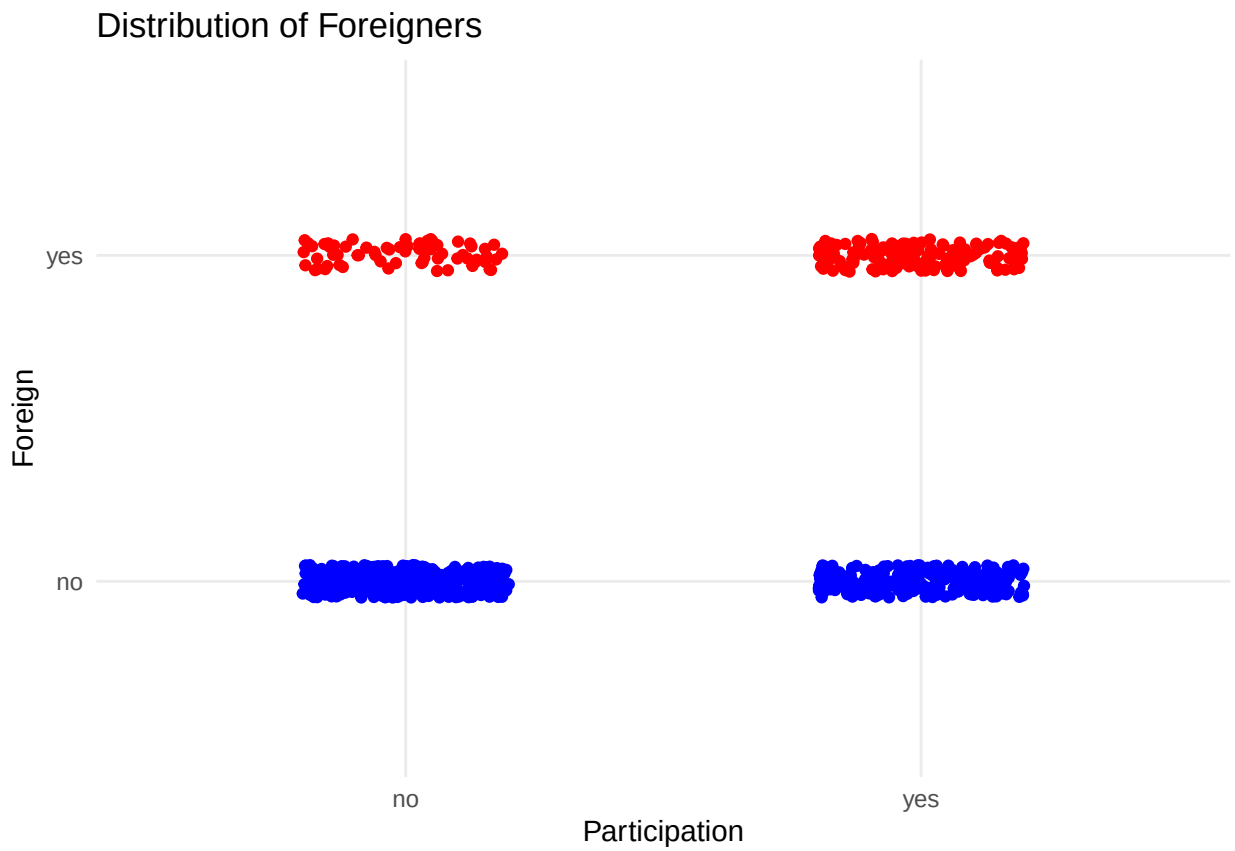
For the FOREIGN variable: Histogram/Fitted Distribution: None because binary variable

Boxplot: None because binary variable

Scatterplot: Graphed it against the other variables

Statistical Summary: None because binary variable

```
#VARIABLE "participation"

#Histogram/Fitted Distribution
#None because binary variable

#Boxplot
#None because binary variable

#Scatterplot
ggplot(SwissLabor, aes(x = participation, y = foreign, color = foreign)) +
  geom_jitter(width = 0.2, height = 0.05) +
  labs(title = "Distribution of Foreigners",
       x = "Participation",
       y = "Foreign") +
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Foreign", "Foreign")) +
  theme_minimal() +
  theme(legend.position = "none")
```



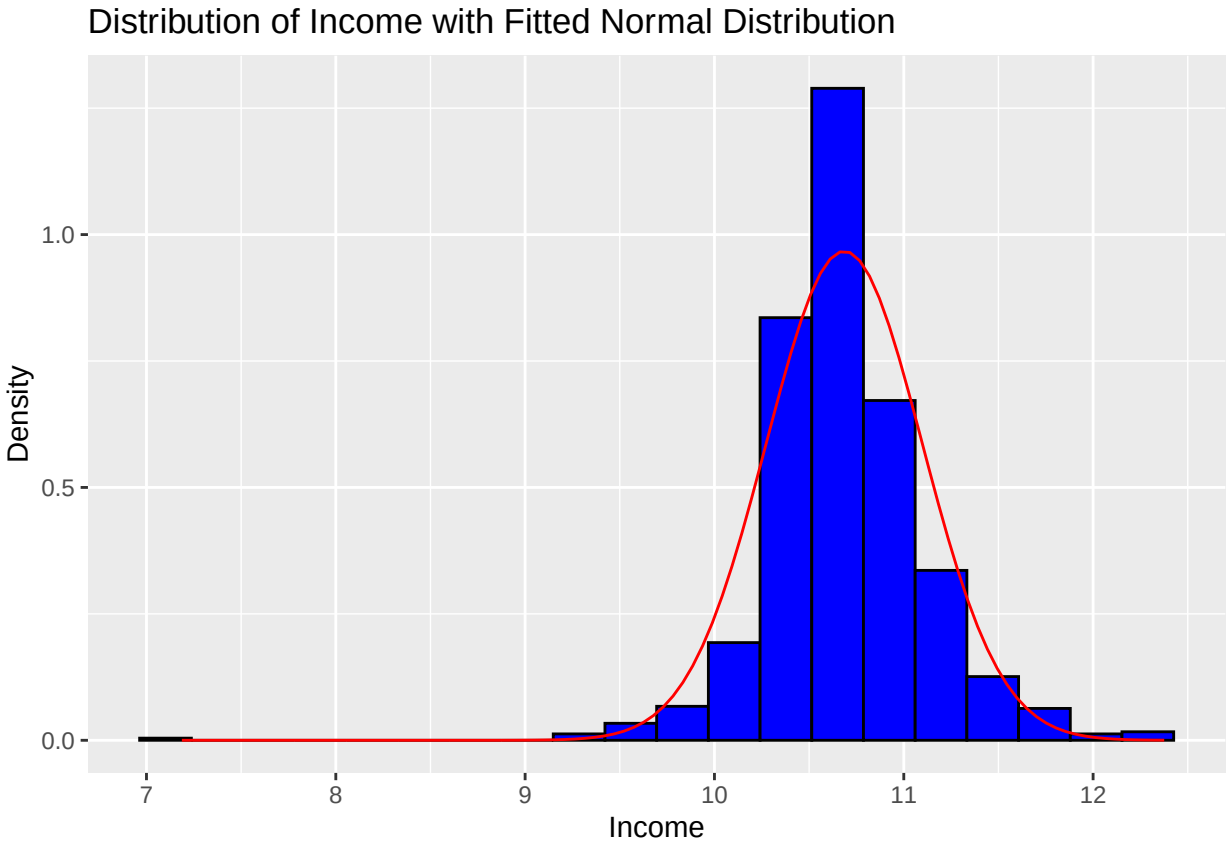

```
#Statistical Summary
#None because binary variable
```

For our participation variable, we don't have a histogram/fitted distribution nor do we have a boxplot because it is a binary variable. We also don't show the statistical summary because it is a binary variable as well.

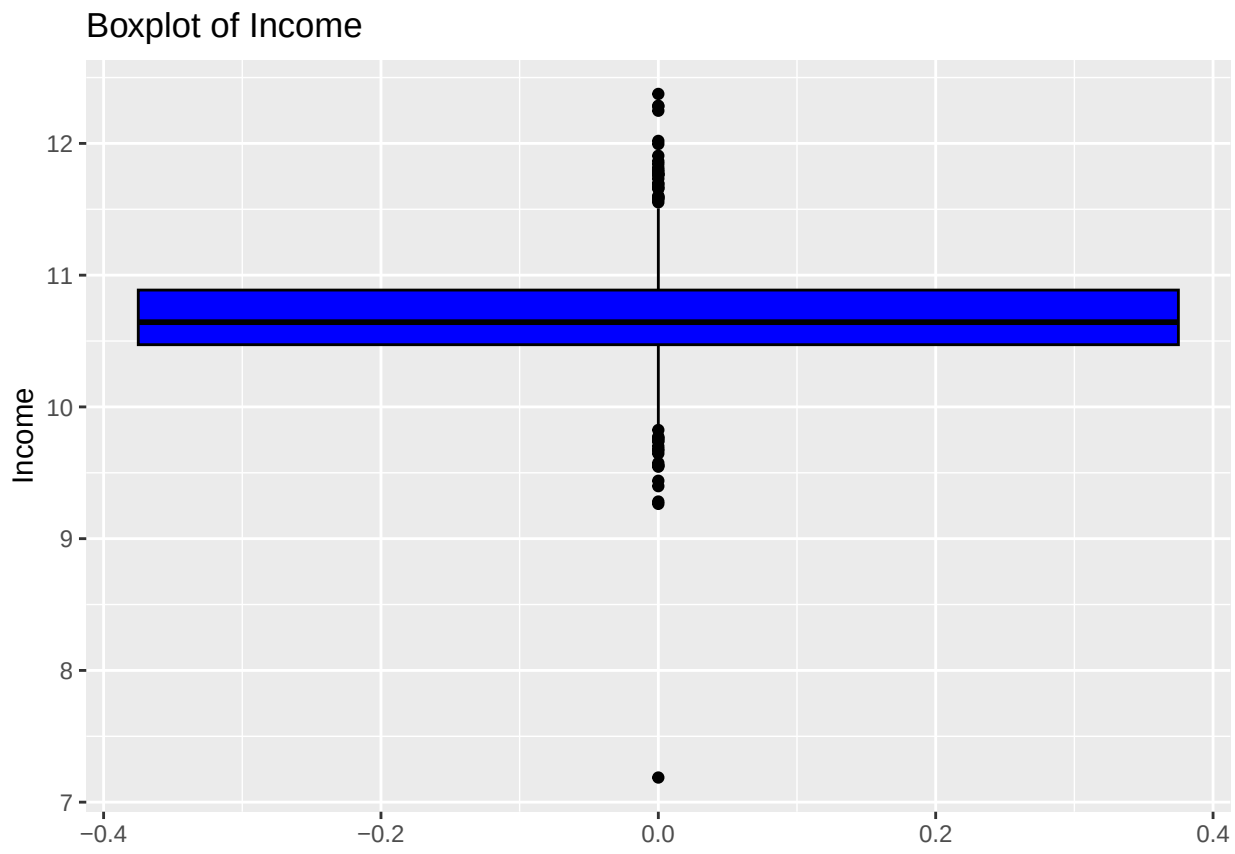
Based on the scatterplot there are more non-foreigners who live in Switzerland than Swiss people because the density of the points are more heavily concentrated in the "no"/blue section.

```
#VARIABLE "income"

#Histogram/Fitted Distribution
ggplot(SwissLabor, aes(x = income)) +
  geom_histogram(aes(y = ..density..), fill = "blue", color = "black", bins = 20) +
  stat_function(fun = dnorm, args = list(mean = mean(SwissLabor$income), sd = sd(SwissLabor$income)), co
  labs(title = "Distribution of Income with Fitted Normal Distribution",
       x = "Income",
       y = "Density")
```

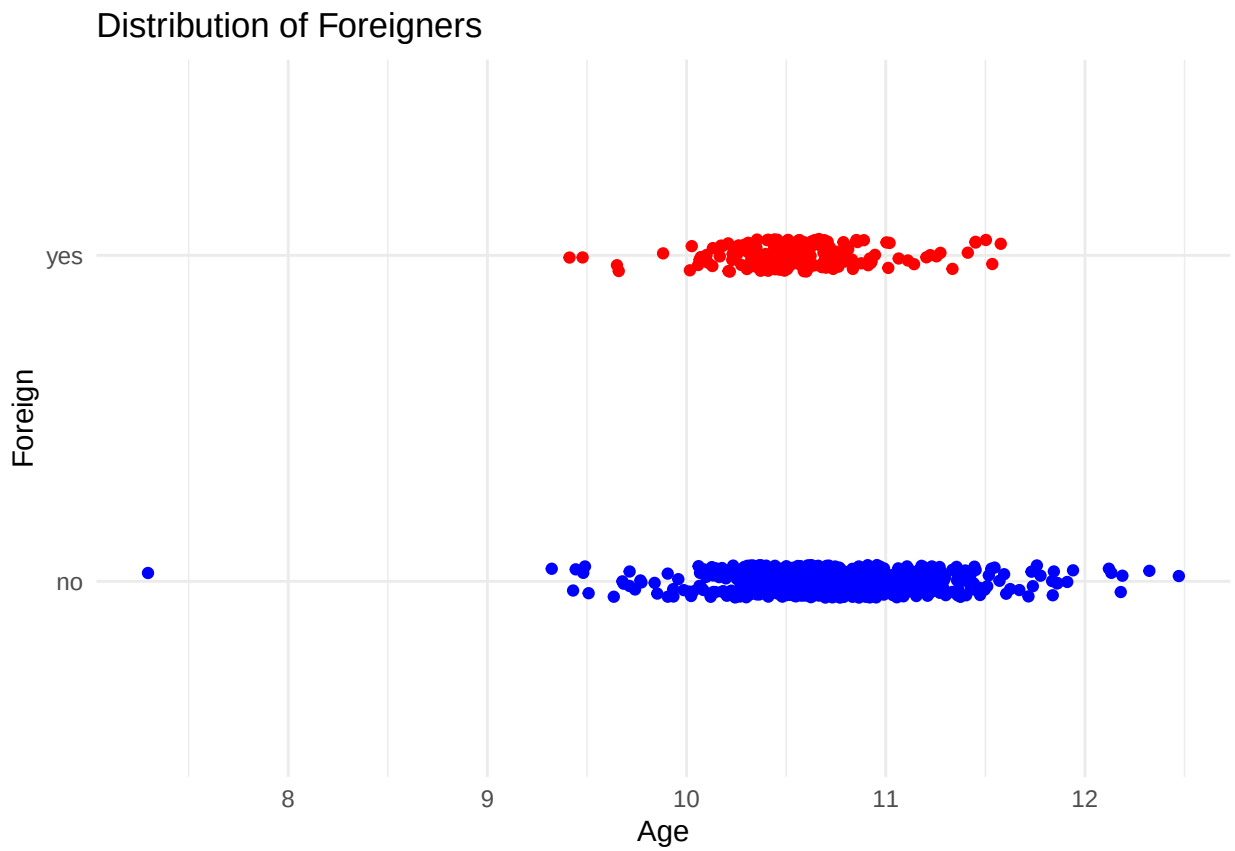
```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
#Boxplot
ggplot(SwissLabor, aes(y = income)) +
  geom_boxplot(fill = "blue", color = "black") +
  labs(title = "Boxplot of Income",
       y = "Income")
```




```
#Scatterplot
ggplot(SwissLabor, aes(x = income, y = foreign, color = foreign)) +
  geom_jitter(width = 0.2, height = 0.05) +
  labs(title = "Distribution of Foreigners",
       x = "Age",
       y = "Foreign") +
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Foreign", "Foreign")) +
  theme_minimal() +
  theme(legend.position = "none")
```

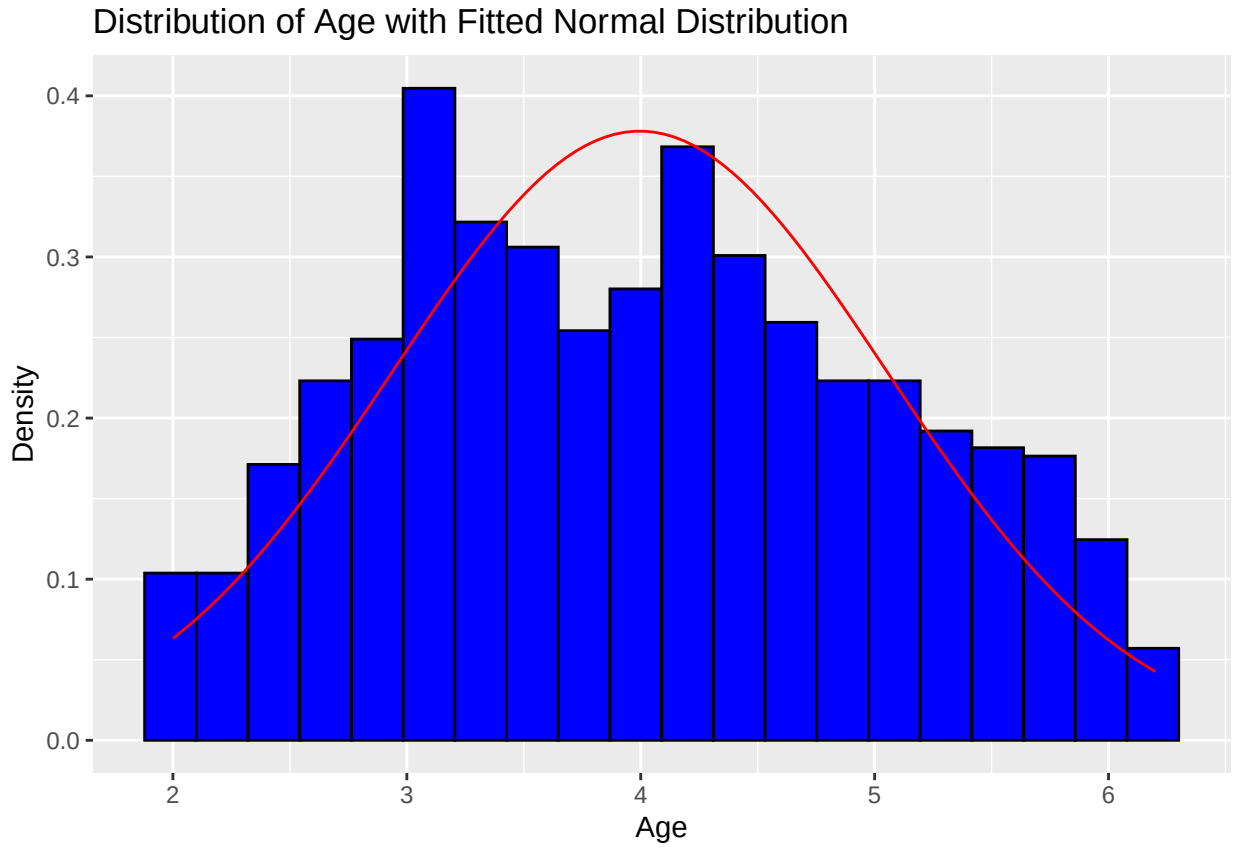



```
#Statistical Summary
summary(SwissLabor$income)
```

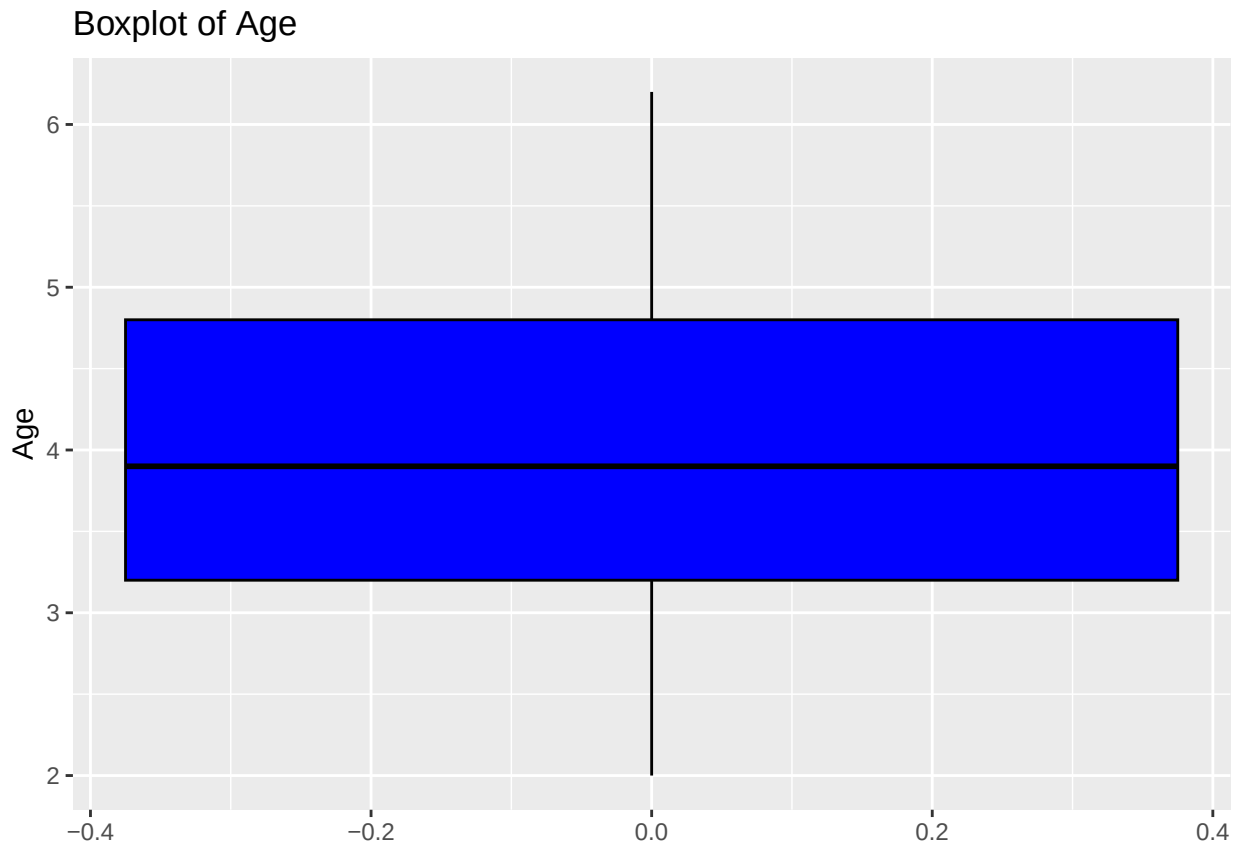
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.187  10.472  10.643  10.686  10.887  12.376
```

The histogram with fitted distribution above shows that the mean/median are both around 10.5 and the bell curve is slightly left-tailed. The boxplot shows that there really aren't any outliers and most of the data is centered around the 10-11 income range. The scatterplot demonstrates that once again there are more non-foreigners than Swiss people; however, in this plot it also shows that the age range of non-foreigners is larger and more diverse than for foreign people. The statistical summary reinforces what the graphs told us in that the mean/median are 10.686 and 10.643 respectively.

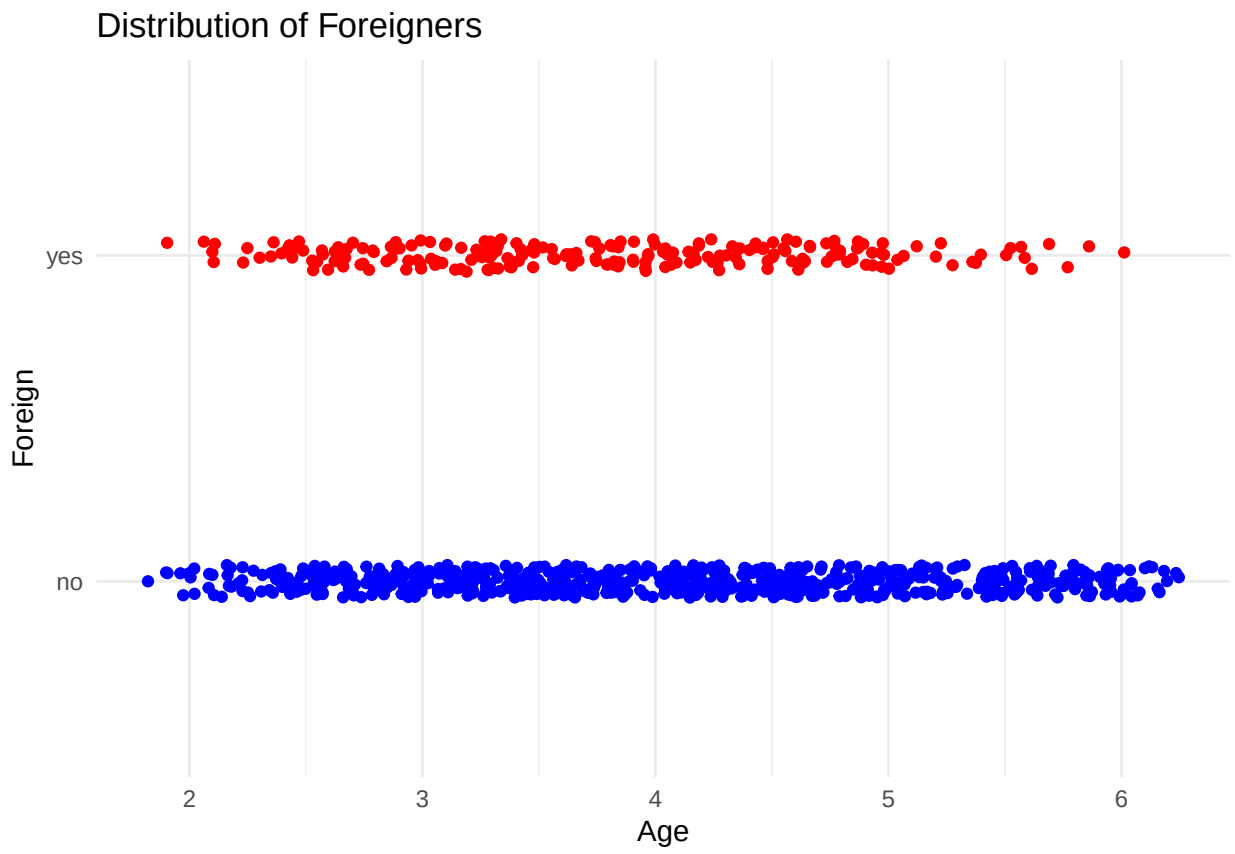
```
#Variable "age"

#Histogram/Fitted Distribution
ggplot(SwissLabor, aes(x = age)) +
  geom_histogram(aes(y = ..density..), fill = "blue", color = "black", bins = 20) +
  stat_function(fun = dnorm, args = list(mean = mean(SwissLabor$age), sd = sd(SwissLabor$age)), color =
  labs(title = "Distribution of Age with Fitted Normal Distribution",
       x = "Age",
       y = "Density")
```




```
#Boxplot
ggplot(SwissLabor, aes(y = age)) +
  geom_boxplot(fill = "blue", color = "black") +
  labs(title = "Boxplot of Age",
       y = "Age")
```




```
#Scatterplot
ggplot(SwissLabor, aes(x = age, y = foreign, color = foreign)) +
  geom_jitter(width = 0.2, height = 0.05) +
  labs(title = "Distribution of Foreigners",
       x = "Age",
       y = "Foreign") +
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Foreign", "Foreign")) +
  theme_minimal() +
  theme(legend.position = "none")
```

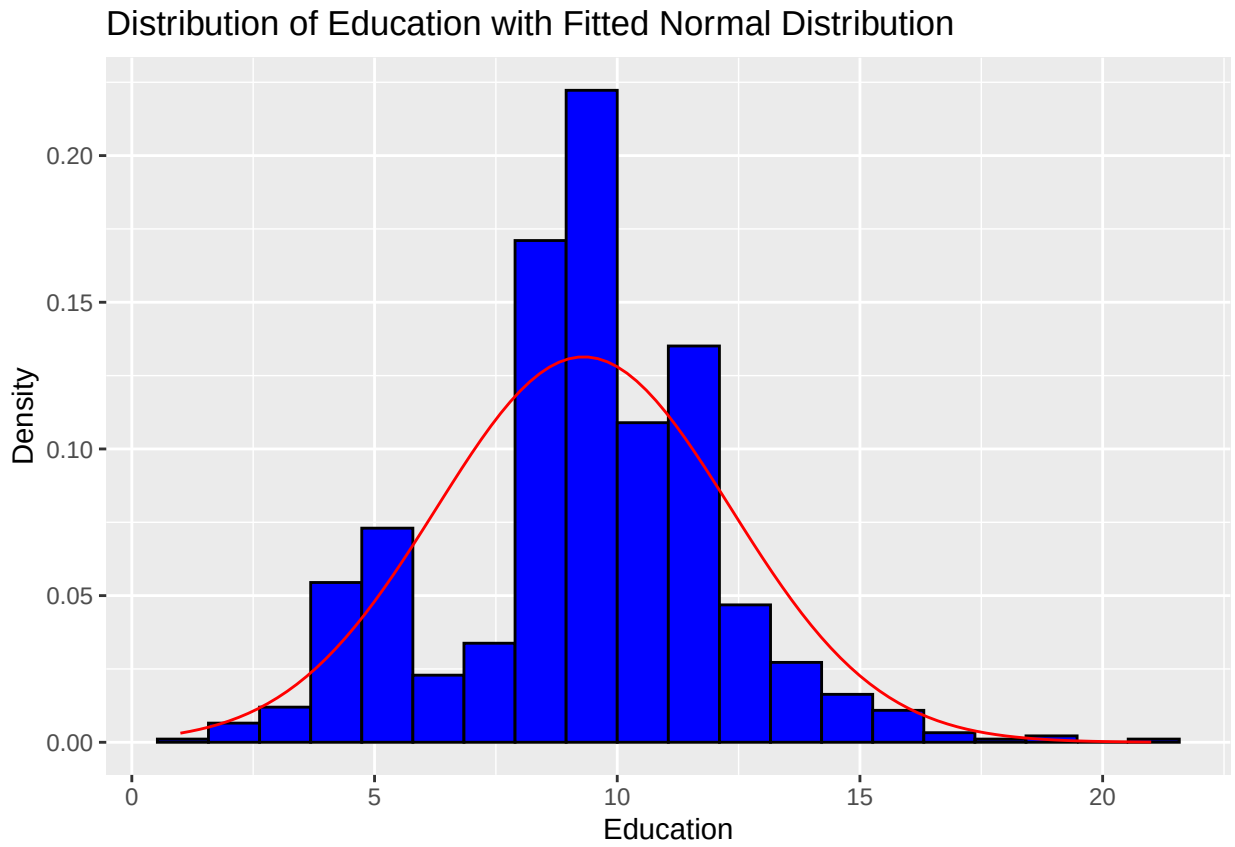



```
#Statistical Summary
summary(SwissLabor$age)
```

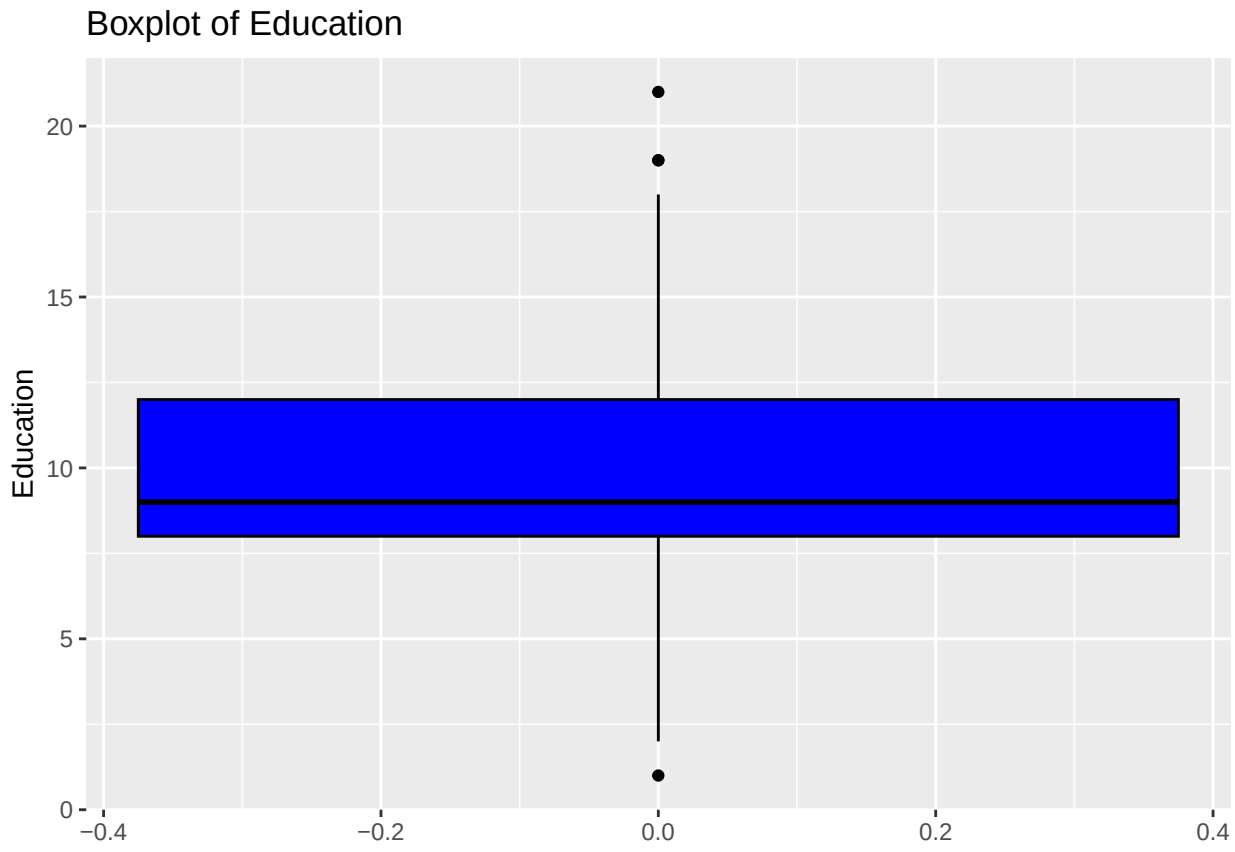
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.200   3.900   3.996   4.800   6.200
```

The histogram with fitted distribution is relatively symmetrical and shows the mean/median is about age 4. The boxplot demonstrates that there aren't any outliers and like the histogram shows the mean/median is at about age 4. The scatterplot illustrates that both non-foreigner and foreigner age distribution ranges are very similar (ranging from ages 2 to 6). The statistical summary supports the different plots and graphs since the mean/median is 3.996 and 3.900 respectively and the min and max are at 2 and 6 respectively.

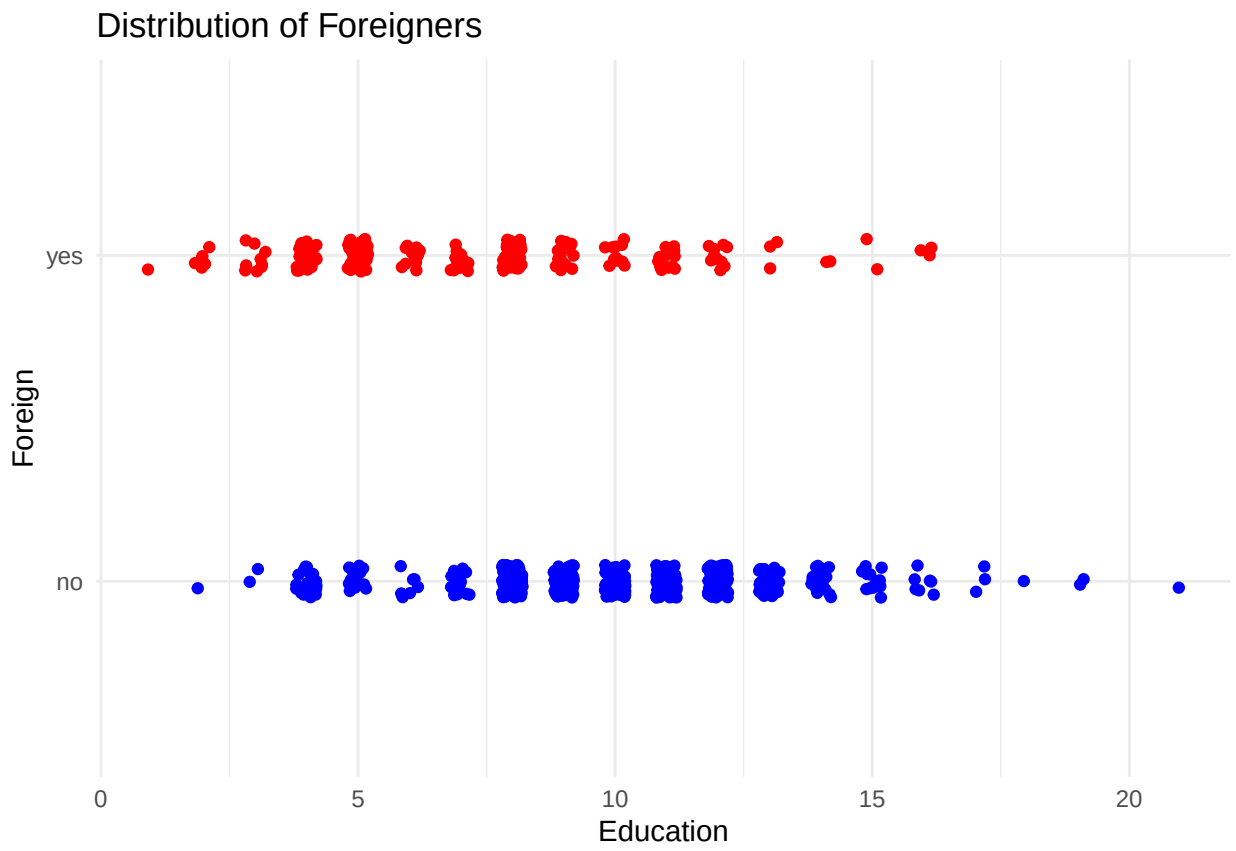
```
#VARIABLE "education"

#Histogram/Fitted Distribution
ggplot(SwissLabor, aes(x = education)) +
  geom_histogram(aes(y = ..density..), fill = "blue", color = "black", bins = 20) +
  stat_function(fun = dnorm, args = list(mean = mean(SwissLabor$education), sd = sd(SwissLabor$educatio
  labs(title = "Distribution of Education with Fitted Normal Distribution",
       x = "Education",
       y = "Density")
```




```
#Boxplot
ggplot(SwissLabor, aes(y = education)) +
  geom_boxplot(fill = "blue", color = "black") +
  labs(title = "Boxplot of Education",
       y = "Education")
```




```
#Scatterplot
ggplot(SwissLabor, aes(x = education, y = foreign, color = foreign)) +
  geom_jitter(width = 0.2, height = 0.05) +
  labs(title = "Distribution of Foreigners",
       x = "Education",
       y = "Foreign") +
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Foreign", "Foreign")) +
  theme_minimal() +
  theme(legend.position = "none")
```

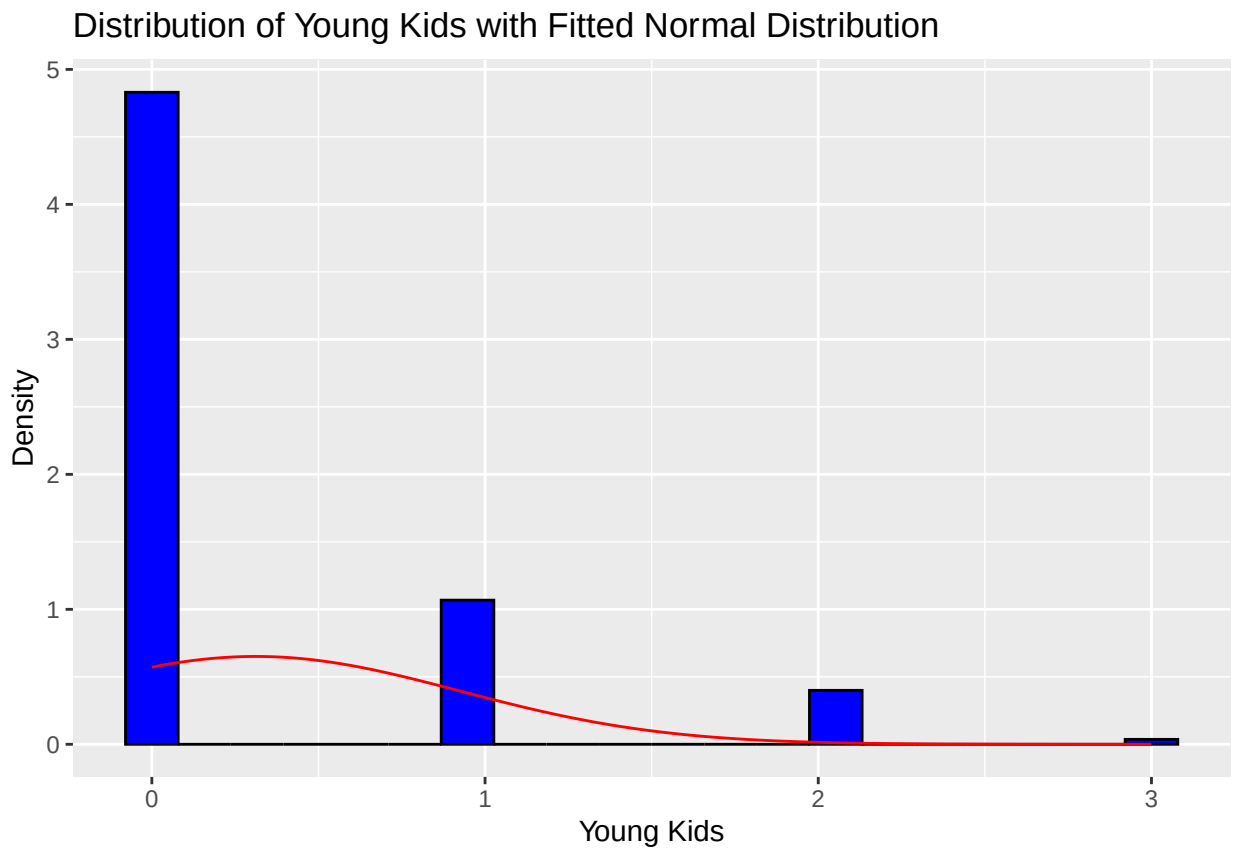



```
#Statistical Summary
summary(SwissLabor$education)
```

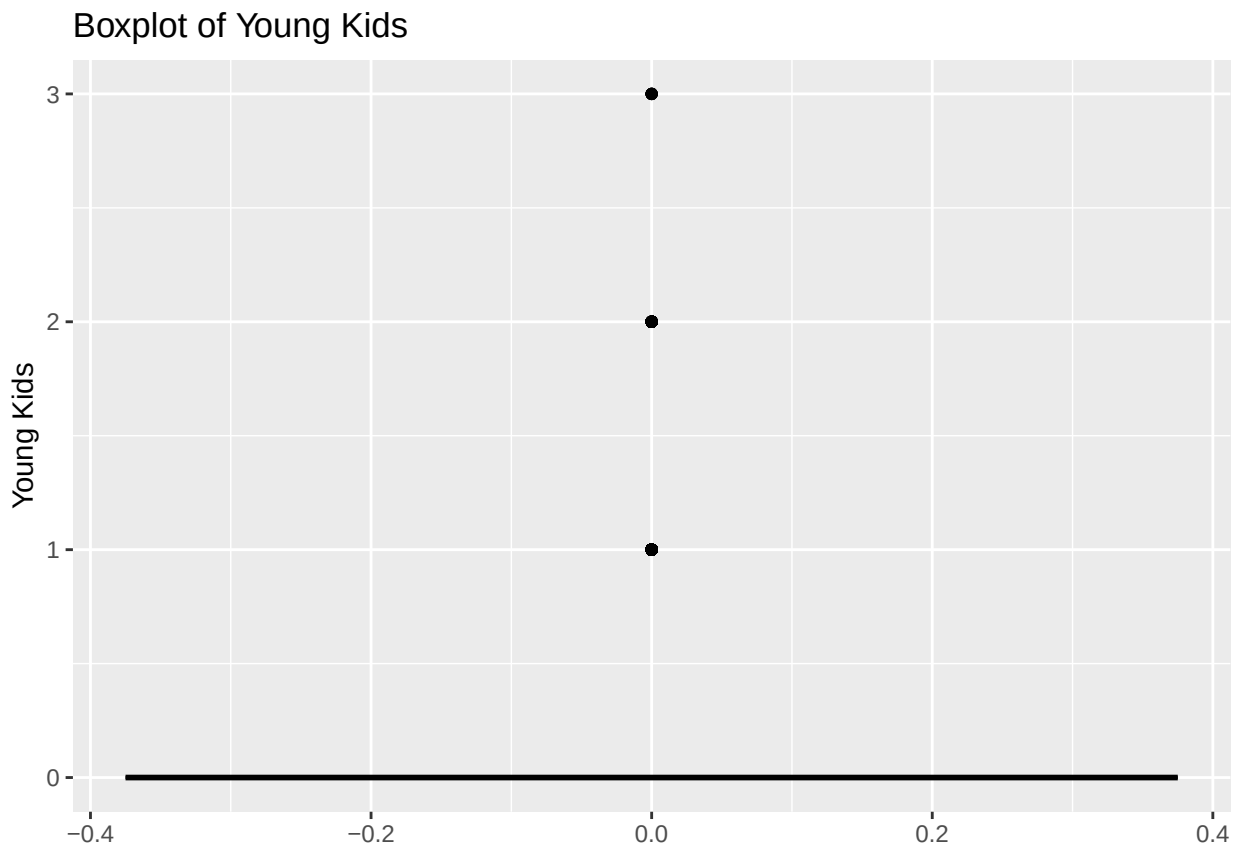
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   8.000   9.000   9.307  12.000  21.000
```

The histogram with fitted distribution looks symmetrical based on the fitted line; however, the columns for years of education at years 6 and 7 are much lower than the fitted distribution. The boxplot demonstrates that the mean/median is around 9-10 years of formal education with very few outliers. The scatterplot shows the average years of education for non-foreigners is higher than the average years of education for foreigners. The statistical summary once again supports the interpretations of the graphs because the mean/median respectively is 9.307 and 9.

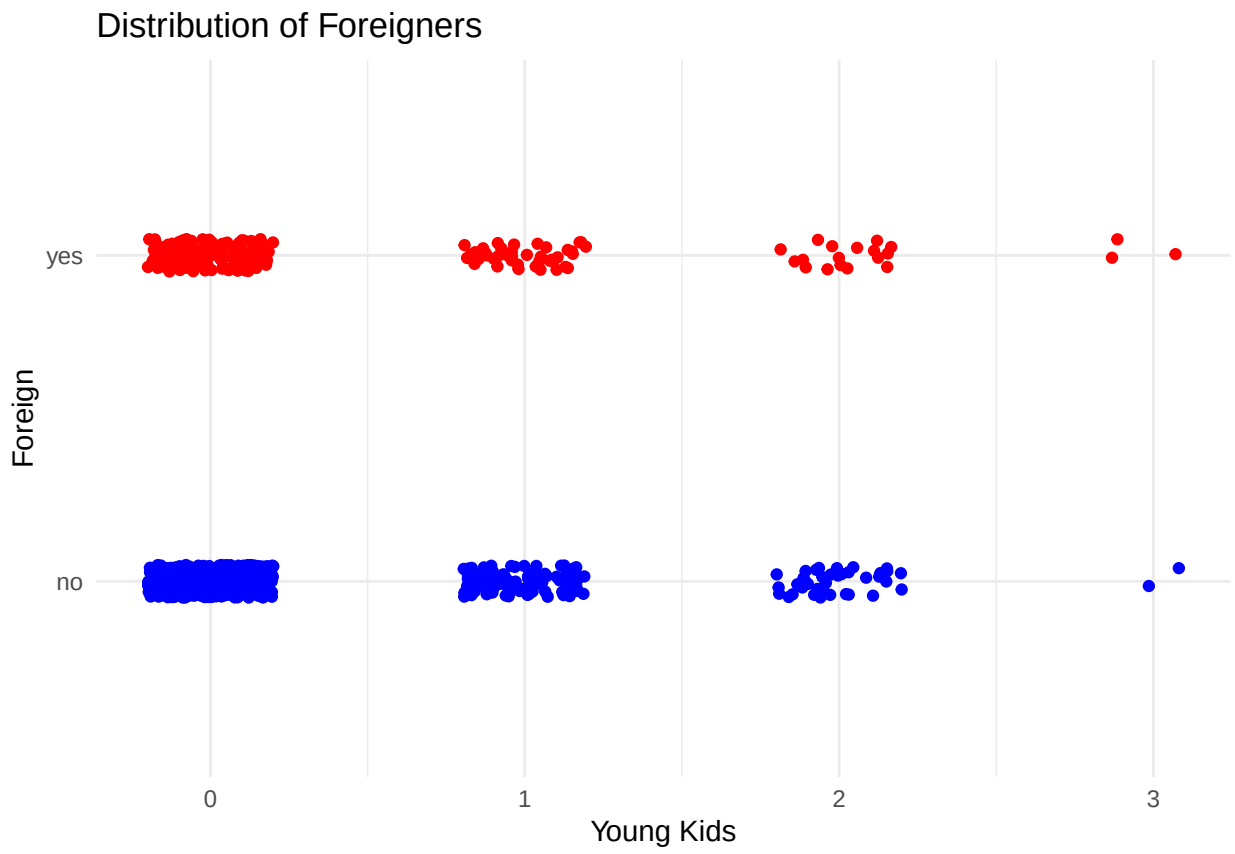
```
#VARIABLE "youngkids"

#Histogram/Fitted Distribution
ggplot(SwissLabor, aes(x = youngkids)) +
  geom_histogram(aes(y = ..density..), fill = "blue", color = "black", bins = 20) +
  stat_function(fun = dnorm, args = list(mean = mean(SwissLabor$youngkids), sd = sd(SwissLabor$youngkids
  labs(title = "Distribution of Young Kids with Fitted Normal Distribution",
       x = "Young Kids",
       y = "Density")
```




```
#Boxplot
ggplot(SwissLabor, aes(y = youngkids)) +
  geom_boxplot(fill = "blue", color = "black") +
  labs(title = "Boxplot of Young Kids",
       y = "Young Kids")
```




```
#Scatterplot
ggplot(SwissLabor, aes(x = youngkids, y = foreign, color = foreign)) +
  geom_jitter(width = 0.2, height = 0.05) +
  labs(title = "Distribution of Foreigners",
       x = "Young Kids",
       y = "Foreign") +
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Foreign", "Foreign")) +
  theme_minimal() +
  theme(legend.position = "none")
```

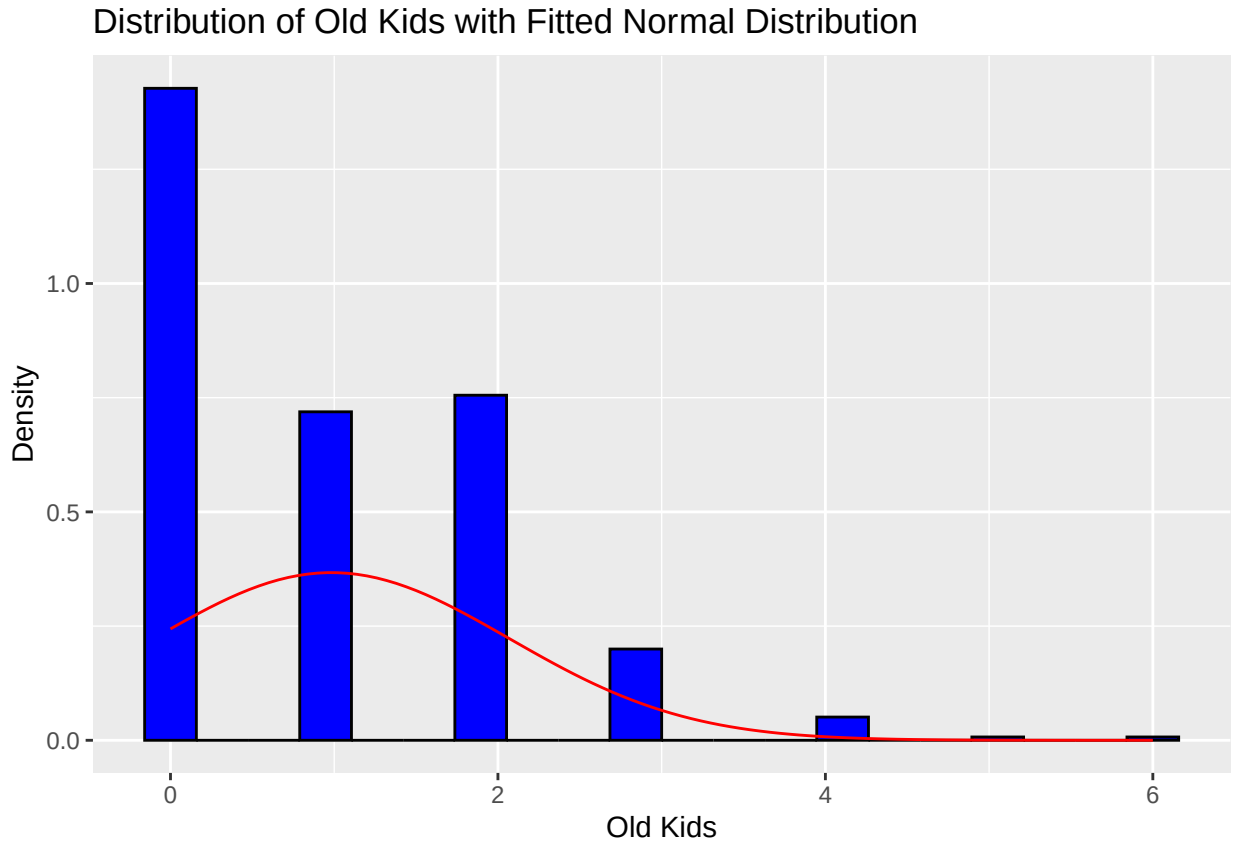



```
#Statistical Summary
summary(SwissLabor$youngkids)
```

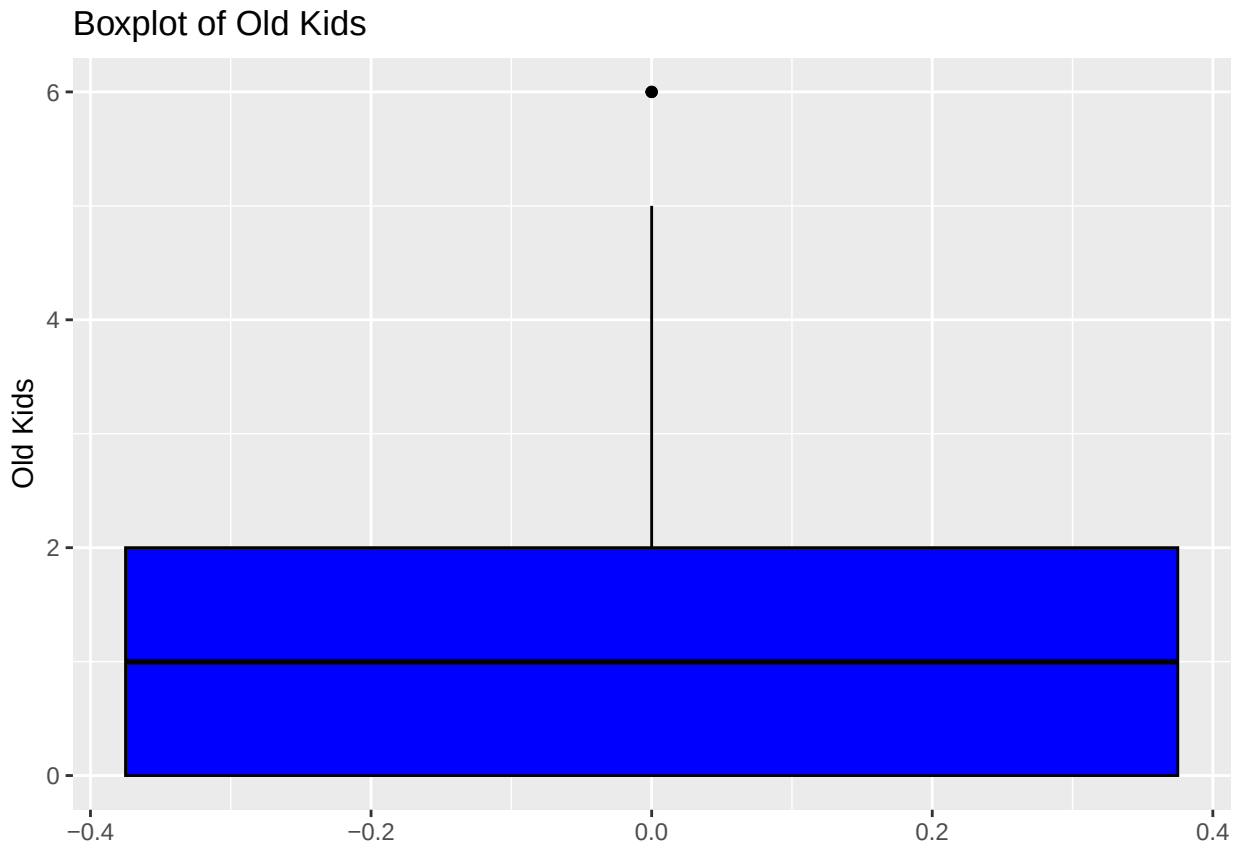
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3119  0.0000  3.0000
```

The histogram with fitted distrbution is right-tailed and this is also shown with the height of the bars. On the left the 0 young kids bar has the highest density by far and it gradually decreases as the number of young kids increases. The boxplot is not evenly distributed as the min and mean/median are all appear to be at 0. This boxplot is also not symmetrical at all. The scatterplot's distribution of the range of number of young kids for non-foreigners and foreigners looks to be very similar. The statistical summary shows that the interpretations of the graphs were correct as the min, 1st quartile, and median are all 0 while the mean is slighly above at about 0.3

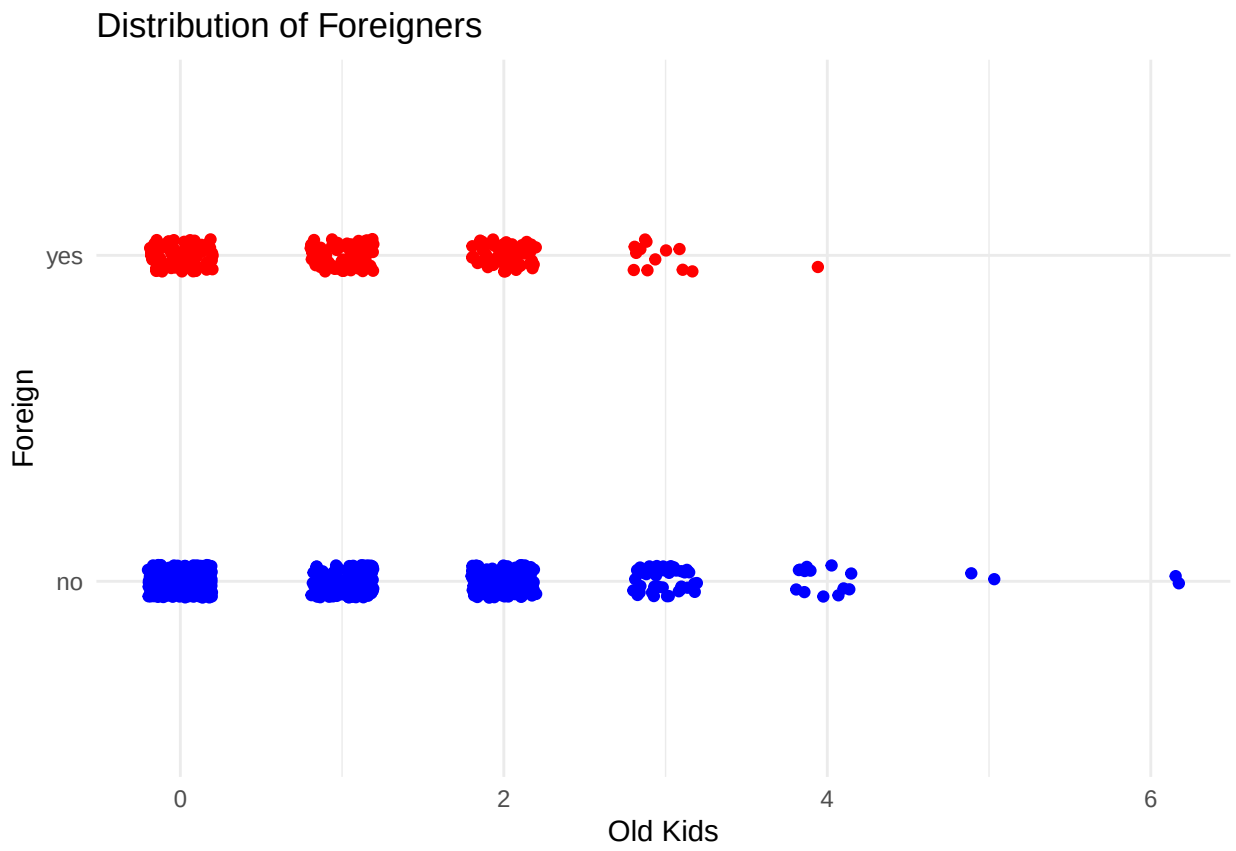
```
#VARIABLE "oldkids"

#Histogram/Fitted Distribution
ggplot(SwissLabor, aes(x = oldkids)) +
  geom_histogram(aes(y = ..density..), fill = "blue", color = "black", bins = 20) +
  stat_function(fun = dnorm, args = list(mean = mean(SwissLabor$oldkids), sd = sd(SwissLabor$oldkids)),
  labs(title = "Distribution of Old Kids with Fitted Normal Distribution",
       x = "Old Kids",
       y = "Density")
```




```
#Boxplot
ggplot(SwissLabor, aes(y = oldkids)) +
  geom_boxplot(fill = "blue", color = "black") +
  labs(title = "Boxplot of Old Kids",
       y = "Old Kids")
```




```
#Scatterplot
ggplot(SwissLabor, aes(x = oldkids, y = foreign, color = foreign)) +
  geom_jitter(width = 0.2, height = 0.05) +
  labs(title = "Distribution of Foreigners",
       x = "Old Kids",
       y = "Foreign") +
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Foreign", "Foreign")) +
  theme_minimal() +
  theme(legend.position = "none")
```




```
#Statistical Summary
summary(SwissLabor$oldkids)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.9828  2.0000  6.0000
```

The histogram with fitted distribution for old kids is similar to the one for young kids in that it is also right-skewed with the first bar of 0 being much larger than the others. The boxplot shows that the mean/median is about 1. This boxplot is also not symmetrical at all. The scatterplot shows non-foreigners generally have more old kids than foreigners. The statistical summary proves the interpretations above are correct, as the mean/median are 0.9828 and 1 respectively and the min is 0.

```
library(dplyr)
```

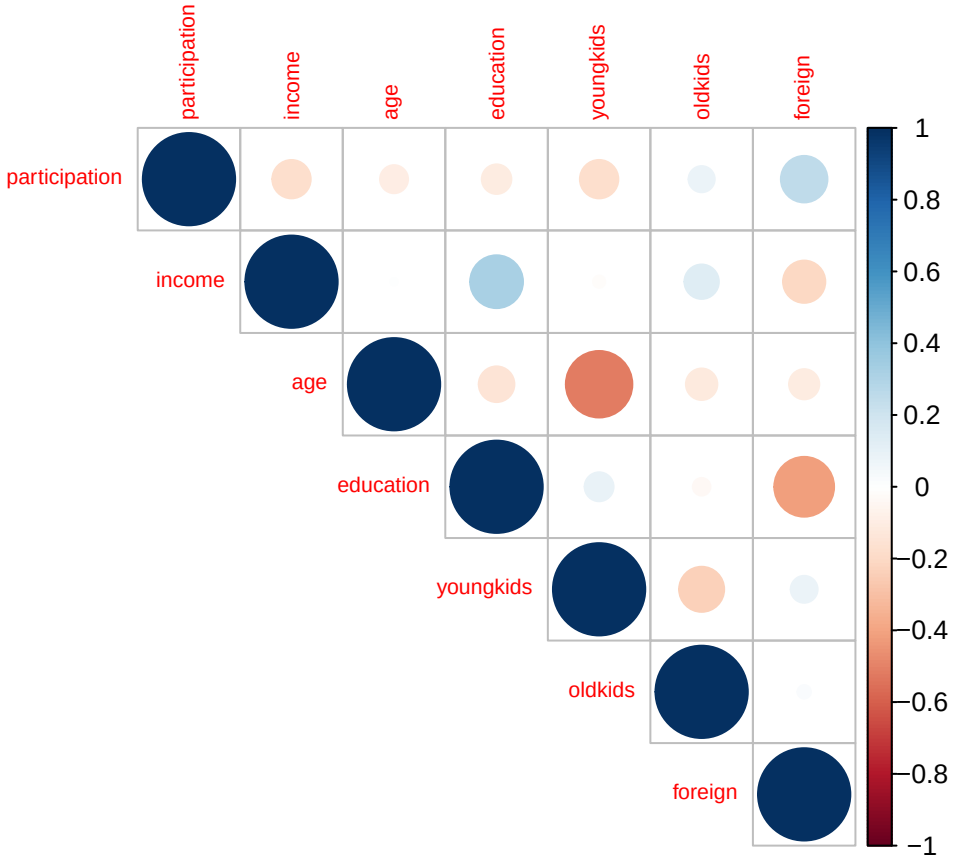
```
## Warning: package 'dplyr' was built under R version 4.1.2
```

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

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:plm':
##
##     between, lag, lead

## The following object is masked from 'package:car':
##
##     recode

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

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

```r
#Correlation Plot
#I combined all of the columns into one plot
# Mutate the participation and foreign columns to 1 and 0 instead of yes and no
SwissLabor <- SwissLabor %>%
      mutate(participation = ifelse(participation == "no",0,1))
SwissLabor <- SwissLabor %>%
      mutate(foreign = ifelse(foreign == "no",0,1))

correlation_matrix <- cor(SwissLabor[sapply(SwissLabor, is.numeric)])
corrplot(correlation_matrix, method = "circle", type = "upper", tl.cex = 0.7)
```



The correlation plot shows that the two variables that are most negatively correlated with one another are age and young kids. The two variables that are the most positively correlated with one another are income and education. Otherwise, we see pretty reasonable correlations across the board except with age and income, young kids and income, and foreign and old kings where the correlations are barely noticeable.

Make models

```
#Linear Probability model

linear_prob <- lm(foreign ~., data = SwissLabor)

summary(linear_prob)
```

```
##
## Call:
## lm(formula = foreign ~ ., data = SwissLabor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7672 -0.2582 -0.1200  0.2343  1.3259
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.220398   0.352404   3.463  0.00056 ***
## participation  0.189092   0.026949   7.017 4.59e-12 ***
## income        -0.038921   0.033476  -1.163  0.24528
## age           -0.033706   0.015220  -2.215  0.02705 *
## education     -0.058185   0.004512 -12.895  < 2e-16 ***
## youngkids      0.086722   0.026729   3.245  0.00122 **
## oldkids        0.005514   0.012752   0.432  0.66554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3742 on 865 degrees of freedom
## Multiple R-squared:  0.2545, Adjusted R-squared:  0.2493
## F-statistic: 49.22 on 6 and 865 DF,  p-value: < 2.2e-16
```

As we can see from the summary, age, years of education, number of young kids, and whether they participated in the workforce is a good indicator of whether or not they are a foreigner.

To interpret each coefficient, for every unit increase in the logarithm of nonlabor income (income is log of nonlabor income), the probability of them being a foreigner decreases by .0389. For every unit increase in age in decades, their probability of being a foreigner decreases by .0337. For every extra year of education, their probability of being a foreigner drops by .058185. For every extra child under 7 they have, their probability of being a foreigner increases by .0867. For every extra child above 7 they have, their probability of being a foreigner increases by .0055. Finally, if they participated in the workforce, their probability of being a foreigner increases by .189.

```
# Look at the results of the model in a table, and calculate the mean error.
lin_prob_predictions <- ifelse(linear_prob$fitted.values > 0.5, 1, 0)

table(lin_prob_predictions, SwissLabor$foreign)
```

```
##
## lin_prob_predictions   0   1
##                    0 633 115
##                    1  23 101
```

```
mean(lin_prob_predictions != SwissLabor$foreign)
```

```
## [1] 0.1582569
```

When looking at the results, we see that most of them are true negatives. This is good because it seems that we can correctly identify which people are not foreigners. However, when it comes to identifying foreigners, we could do a better job. There were 115 people that we classified as not being foreigners that actually did end up being foreigners, and we only classified 101 foreigners correctly. Only 23 people that we identified as foreigners turned out to not be foreigners, which is good. Overall, this model seemed to stick to the commonality of the data, which was that most people weren't foreigners. We have approximately a 16% error rate, which is very good. This model could use some work, but this is a very good starting point. We could also try a version that doesn't contain insignificant predictors to see if that makes a difference.

Before we cut predictors, one important thing to note is the effectiveness of a naive classifier. While our model does looks good for now, it is important to note that most values in the dataset are non-foreigner and most predictions are non-foreigner. We have to account for the possibility that we may get better results if we just pick everyone to be a non-foreigner, in which case it would mean that our model is not good.

```
# Look at how accurate a naive classifier would be
length(SwissLabor$foreign[SwissLabor$foreign == "0"]) / length(SwissLabor$foreign)
```

```
## [1] 0.7522936
```

Since our model is 84% accurate and a naive classifier is 75% accurate, this shows that our model is an improvement over the bare minimum model, and is good to build off of.

```
# Cut down on predictors and rerun the linear probability model and calculate predictions and error sta
linear_prob_edited <- lm(foreign ~ age + education + youngkids + participation, data = SwissLabor)

summary(linear_prob_edited)
```

```
##
## Call:
## lm(formula = foreign ~ age + education + youngkids + participation,
##     data = SwissLabor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7679 -0.2559 -0.1221  0.2181  1.3192
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.832905   0.083235  10.007  < 2e-16 ***
## age           -0.035664   0.014536  -2.454 0.014342 *
## education     -0.059970   0.004252 -14.104  < 2e-16 ***
## youngkids      0.084611   0.025059   3.377 0.000767 ***
## participation  0.193677   0.026614   7.277 7.64e-13 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3741 on 867 degrees of freedom
## Multiple R-squared:  0.2533, Adjusted R-squared:  0.2498
## F-statistic: 73.52 on 4 and 867 DF,  p-value: < 2.2e-16
```

```
lin_prob_predictions_edited <- ifelse(linear_prob_edited$fitted.values > 0.5, 1, 0)

table(lin_prob_predictions_edited, SwissLabor$foreign)
```

```
##
## lin_prob_predictions_edited   0   1
##                           0 634 119
##                           1  22  97
```

```
mean(lin_prob_predictions_edited != SwissLabor$foreign)
```

```
## [1] 0.1616972
```

When cutting down on predictors, we see that all the predictors are significant, but our error rate actually increased by a little bit. For now, the best model is the original one with all the predictors

```
#We will now try a probit model
probit <- glm(foreign ~ ., family = binomial(link="probit"), data=SwissLabor)

summary(probit)
```

```
##
## Call:
## glm(formula = foreign ~ ., family = binomial(link = "probit"),
##     data = SwissLabor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9669  -0.6857  -0.4420  -0.0439   3.6132
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.383465   1.564222   2.163  0.03054 *
## participation  0.699812   0.110216   6.349 2.16e-10 ***
## income        -0.199829   0.147572  -1.354  0.17570
## age           -0.140570   0.062523  -2.248  0.02456 *
## education     -0.210940   0.020139 -10.474  < 2e-16 ***
## youngkids      0.295553   0.104733   2.822  0.00477 **
## oldkids        0.006948   0.051559   0.135  0.89281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 976.29  on 871  degrees of freedom
```

```
## Residual deviance: 741.47  on 865  degrees of freedom
## AIC: 755.47
##
## Number of Fisher Scoring iterations: 5
```

For those who participated in the workforce, the z-score increases by approximately .7. For a one unit increase in log of nonlabor income, the z-score decreases by approximately .2. For a one unit increase in age in decades, the z-score decreases by .14. For a one unit increase in years of education, the z-score decreases by .21. For every extra child under the age of 7, the z-score increases by approximately .3. For every child above the age of 7, the z-score increases by approximately .007.

Everything except income and old kids is significant

```
# Look at the results from the probit model and error of the model.
probit_pred <- ifelse(probit$fitted.values > 0.5, "1", "0")

table(probit_pred, SwissLabor$foreign)
```

```
##
## probit_pred   0   1
##           0 630 112
##           1  26 104
```

```
mean(probit_pred != SwissLabor$foreign)
```

```
## [1] 0.1582569
```

Similar to last time, we get approximately a 16% error rate (exact same as linear probability!). However, it should be noted that the values are a little different. We correctly identified 3 more foreigners, but this came at the cost of incorrectly identifying 3 more non-foreigners.

```
#Finally, we will try the logit model, and look at the results of the predictions
logit <- glm(foreign ~ ., family = binomial(link="logit"), data=SwissLabor)

summary(probit)
```

```
##
## Call:
## glm(formula = foreign ~ ., family = binomial(link = "probit"),
##     data = SwissLabor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9669  -0.6857  -0.4420  -0.0439   3.6132
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.383465   1.564222   2.163  0.03054 *
## participation  0.699812   0.110216   6.349 2.16e-10 ***
## income        -0.199829   0.147572  -1.354  0.17570
## age           -0.140570   0.062523  -2.248  0.02456 *
## education     -0.210940   0.020139 -10.474  < 2e-16 ***
```

```
## youngkids      0.295553   0.104733   2.822  0.00477 **
## oldkids        0.006948   0.051559   0.135  0.89281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 976.29  on 871  degrees of freedom
## Residual deviance: 741.47  on 865  degrees of freedom
## AIC: 755.47
##
## Number of Fisher Scoring iterations: 5
```

```
logit_pred <- ifelse(logit$fitted.values > 0.5, "1", "0")

table(logit_pred, SwissLabor$foreign)
```

```
##
## logit_pred   0   1
##          0 629 109
##          1  27 107
```

```
mean(logit_pred != SwissLabor$foreign)
```

```
## [1] 0.1559633
```

While it was just barely better, we see that the logit model does a better job with predictions than probit and/or linear probability. This one is at 15.6%, while the others were at 15.8%. This one actually had less accurate non-foreigner classifications but made up for it with even more accurate foreigner-predictions. In terms of the summary, everything except income and oldkids is significant.

```
# Transform values for interpretation
exp(1.25493)
```

```
## [1] 3.507593
```

```
(1 - exp(-.48714)) * 100
```

```
## [1] 38.5619
```

```
(1 - exp(-.30176)) * 100
```

```
## [1] 26.04845
```

```
(1 - exp(-.39936)) * 100
```

```
## [1] 32.92508
```

```
(1 - exp(.54484)) * 100
```

```
## [1] -72.43325
```

```
(1 - exp(.03578)) * 100
```

```
## [1] -3.642781
```

For those that participated in the workforce, their odds of being a foreigner are 3.5 times higher. For every 1 unit increase in income, the odds of being a foreigner decrease by 38.56%. For every 1 unit increase in age, the odds of being a foreigner decreases by 26%. For every 1 unit increase in education, the odds of being a foreigner decreases by 32.93%. For every 1 unit increase in young kids, the odds of being a foreigner increases by 72.43%. For every 1 unit increase in old kids, the odds of being a foreigner increases by 3.64%

```
# Look at the AIC BIC values for the models
AIC(linear_prob)
```

```
## [1] 769.4342
```

```
AIC(probit)
```

```
## [1] 755.4652
```

```
AIC(logit)
```

```
## [1] 740.4748
```

```
BIC(linear_prob)
```

```
## [1] 807.6005
```

```
BIC(probit)
```

```
## [1] 788.8607
```

```
BIC(logit)
```

```
## [1] 773.8703
```

AIC and BIC tests are another good way of seeing which model we should, and in this case, both AIC and BIC support the logit model, just as the error rate did.

Using your preferred model, make 4 different predictions, and comment on their reliability.

```r
# Make the predictions using our logit model and look at reliability based on misclassification rate.
logit_pred_values <- predict(logit, type = "response")

# Prediction 1: Using a threshold of 0.4
logit_pred_1 <- ifelse(logit_pred_values > 0.4, "1", "0")

# Prediction 2: Using a threshold of 0.5
logit_pred_2 <- ifelse(logit_pred_values > 0.5, "1", "0")

# Prediction 3: Using a threshold of 0.6
logit_pred_3 <- ifelse(logit_pred_values > 0.6, "1", "0")

# Prediction 4: Using a threshold of 0.7
logit_pred_4 <- ifelse(logit_pred_values > 0.7, "1", "0")

# Confusion matrix and misclassification rate for Prediction 1
confusion_matrix_1 <- table(logit_pred_1, SwissLabor$foreign)
misclassification_rate_1 <- mean(logit_pred_1 != SwissLabor$foreign)

# Confusion matrix and misclassification rate for Prediction 2
confusion_matrix_2 <- table(logit_pred_2, SwissLabor$foreign)
misclassification_rate_2 <- mean(logit_pred_2 != SwissLabor$foreign)

# Confusion matrix and misclassification rate for Prediction 3
confusion_matrix_3 <- table(logit_pred_3, SwissLabor$foreign)
misclassification_rate_3 <- mean(logit_pred_3 != SwissLabor$foreign)

# Confusion matrix and misclassification rate for Prediction 4
confusion_matrix_4 <- table(logit_pred_4, SwissLabor$foreign)
misclassification_rate_4 <- mean(logit_pred_4 != SwissLabor$foreign)

# Print the confusion matrices and misclassification rates
cat("Prediction 1 (Threshold = 0.4):\n")
```

```
## Prediction 1 (Threshold = 0.4):
```

```r
print(confusion_matrix_1)
```

```
##
## logit_pred_1   0   1
##            0 605  83
##            1  51 133
```

```r
cat("Misclassification Rate:", misclassification_rate_1, "\n\n")
```

```
## Misclassification Rate: 0.1536697
```

```r
cat("Prediction 2 (Threshold = 0.5):\n")
```

```
## Prediction 2 (Threshold = 0.5):
```

```
print(confusion_matrix_2)
```

```
##
## logit_pred_2   0   1
##            0 629 109
##            1  27 107
```

```
cat("Misclassification Rate:", misclassification_rate_2, "\n\n")
```

```
## Misclassification Rate: 0.1559633
```

```
cat("Prediction 3 (Threshold = 0.6):\n")
```

```
## Prediction 3 (Threshold = 0.6):
```

```
print(confusion_matrix_3)
```

```
##
## logit_pred_3   0   1
##            0 639 128
##            1  17  88
```

```
cat("Misclassification Rate:", misclassification_rate_3, "\n\n")
```

```
## Misclassification Rate: 0.1662844
```

```
cat("Prediction 4 (Threshold = 0.7):\n")
```

```
## Prediction 4 (Threshold = 0.7):
```

```
print(confusion_matrix_4)
```

```
##
## logit_pred_4   0   1
##            0 644 154
##            1  12  62
```

```
cat("Misclassification Rate:", misclassification_rate_4, "\n")
```

```
## Misclassification Rate: 0.190367
```

The prediction that returned the best misclassification rate was prediction 1 at a threshold of >0.4 because it had the lowest misclassification rate (0.153). Since all of the predictions return a similar misclassification rate it demonstrates that there is consistency between the different thresholds and altering the thresholds does not hold much significance. Furthermore, since all of the misclassification rates are low, the logit model has good reliability.

Conclusion

So overall, through the panel and qualitative dependent parts of our project, we were able to fit models to the data, and come up with preferred models. First, in the panel data section, we fit a pooled model, fixed effects model, and random effects model to our Grunfeld data and came to the conclusion that our preferred model was the fixed effects model, but just based on firms and not a time effect.

Then, in the qualitative dependent modeling part of our model, we fit a linear probability, probit, and logit model to our SwissLabor data and through testing and analysis, came to the conclusion that our Logit model was our preferred model, and through our predictions we can see that our logit model had fairly good reliability.