Econ 104 Project 3

Ernesto Favela, Sara Charolia, Paul Valeriano, Matthew Craig

2023-03-16

Contents

```
rm(list=ls(all=TRUE))
library(AER)
library(car)
library(coefplot)
library(corrplot)
library(dplyr)
library(ggplot2)
library(gplots)
library(knitr)
library(plm)
library(pROC)
library(stargazer)
```

Panel Data Model

1. Dataset

Our Grunfeld panel dataset includes eleven firms, General Motors, US Steel, General Electric, Chrysler, Atlantic Refining, IBM, Union Oil, Westinghouse, Goodyear, Diamond Match, and American Steel, from 1935 to 1954. The factors we will be taking into account are gross investments, the value of the firm, and capital (stock of PP&E). Using this dataset we are trying to determine whether time or the firm's size has a greater effect on the market value.

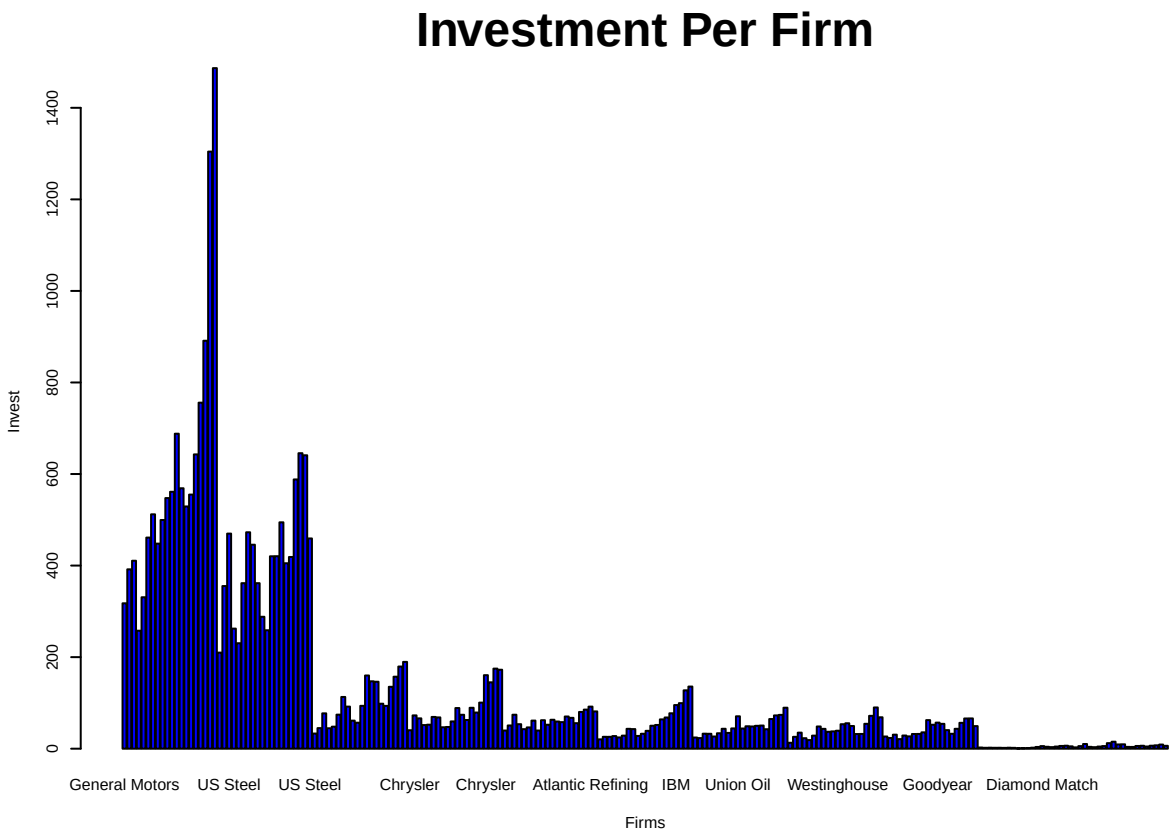
```
data("Grunfeld", package = "AER")
invest <- Grunfeld$invest
value <- Grunfeld$value
capital <- Grunfeld$capital
firm <- Grunfeld$firm
year <- Grunfeld$year
summary(Grunfeld)
```

```
##      invest            value            capital                     firm
##  Min.   :   0.93   Min.   :  30.28   Min.   :   0.8   General Motors   : 20
##  1st Qu.:  27.38   1st Qu.: 160.32   1st Qu.:  67.1   US Steel         : 20
##  Median :  52.37   Median : 404.65   Median : 180.1   General Electric : 20
##  Mean   : 133.31   Mean   : 988.58   Mean   : 257.1   Chrysler         : 20
##  3rd Qu.:  99.78   3rd Qu.:1605.92   3rd Qu.: 344.5   Atlantic Refining: 20
##  Max.   :1486.70   Max.   :6241.70   Max.   :2226.3   IBM              : 20
##                                                       (Other)          :100
##       year
##  Min.   :1935
##  1st Qu.:1940
##  Median :1944
##  Mean   :1944
##  3rd Qu.:1949
##  Max.   :1954
##
```

2. Descriptive Analysis

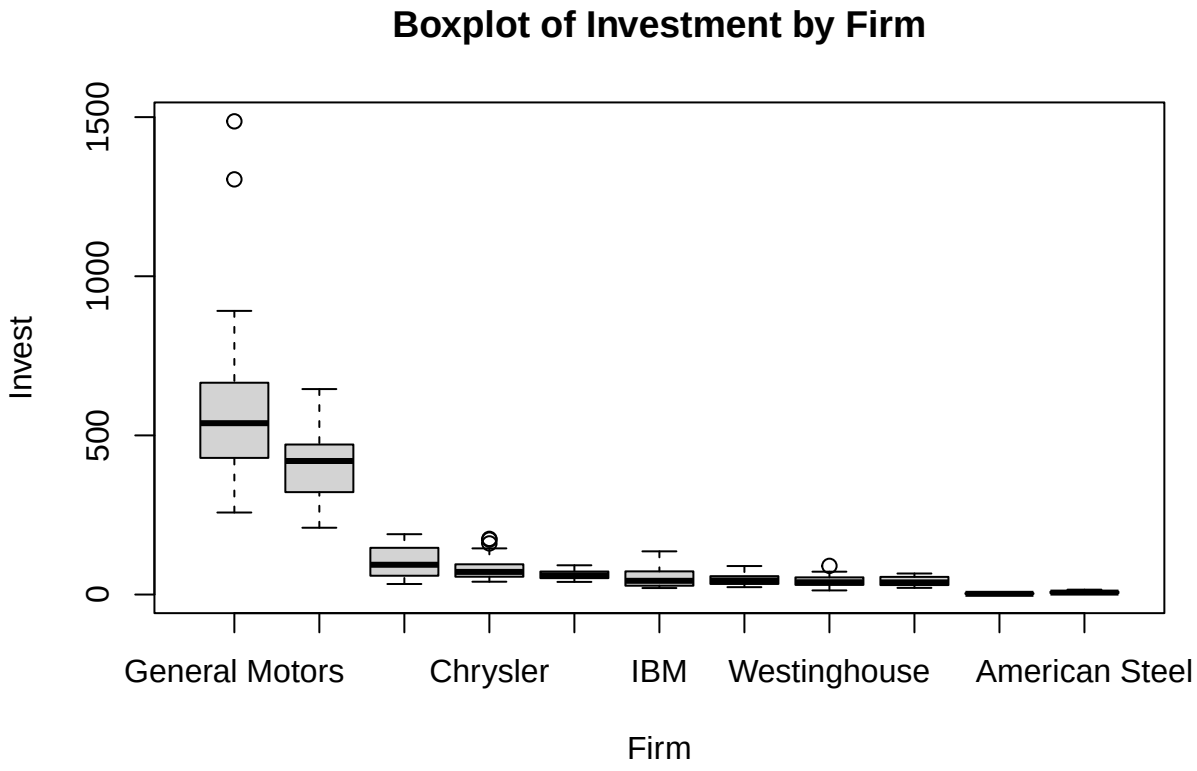
Investment

```
par(cex = 0.5)
par(cex.main = 3)
barplot(height = invest, names = firm, xlab = "Firms", ylab = "Invest",
        main = "Investment Per Firm", col = "blue")
```



The barplot is skewed right for each firm, showing the increasing amount of investment for each firm over the years. We can see that General Motors and US Steel have had the largest exponential growth in investment out of the eleven firms, with scales much larger than the other 9 firms in the dataset.

```
plot(firm, invest, main="Boxplot of Investment by Firm", xlab = "Firm", ylab = "Invest")
```



From the boxplot, we can see that firms with larger investments tend to have a larger variance in the amount invested for that firm. Again, we see that the levels of General Motors and U.S. Steel are much higher than for the other firms in the dataset.

```
hist(invest, main="Histogram of Investment")
```

The long right tail of the histogram indicates that most values fall in the lower end under 200, with a few extremely large outliers

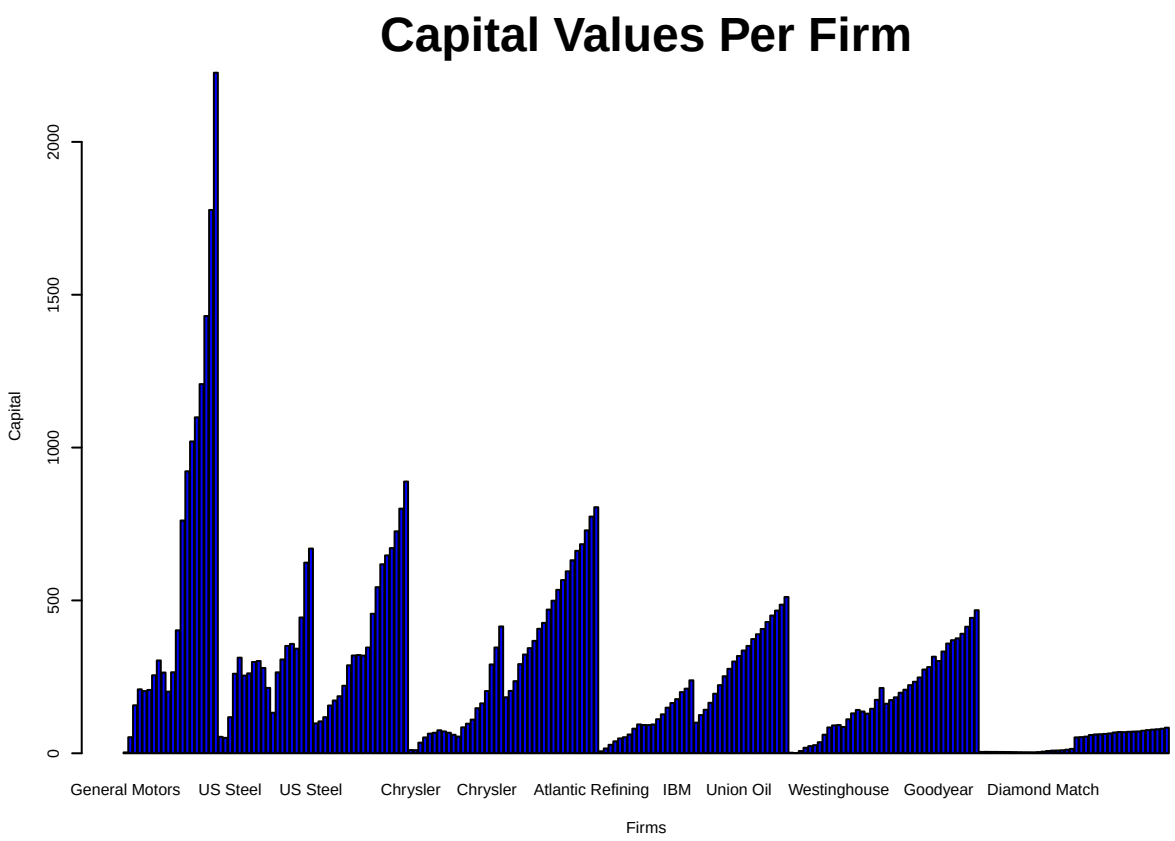
```
# Min, Q1, Median, Q3, Max
fivenum(invest)
```

```
## [1]    0.930   27.230   52.365  100.075 1486.700
```

The five-number summary reinforces the long right tail in the data. The IQR is about 73, while the maximum value is 1386 above Q3.

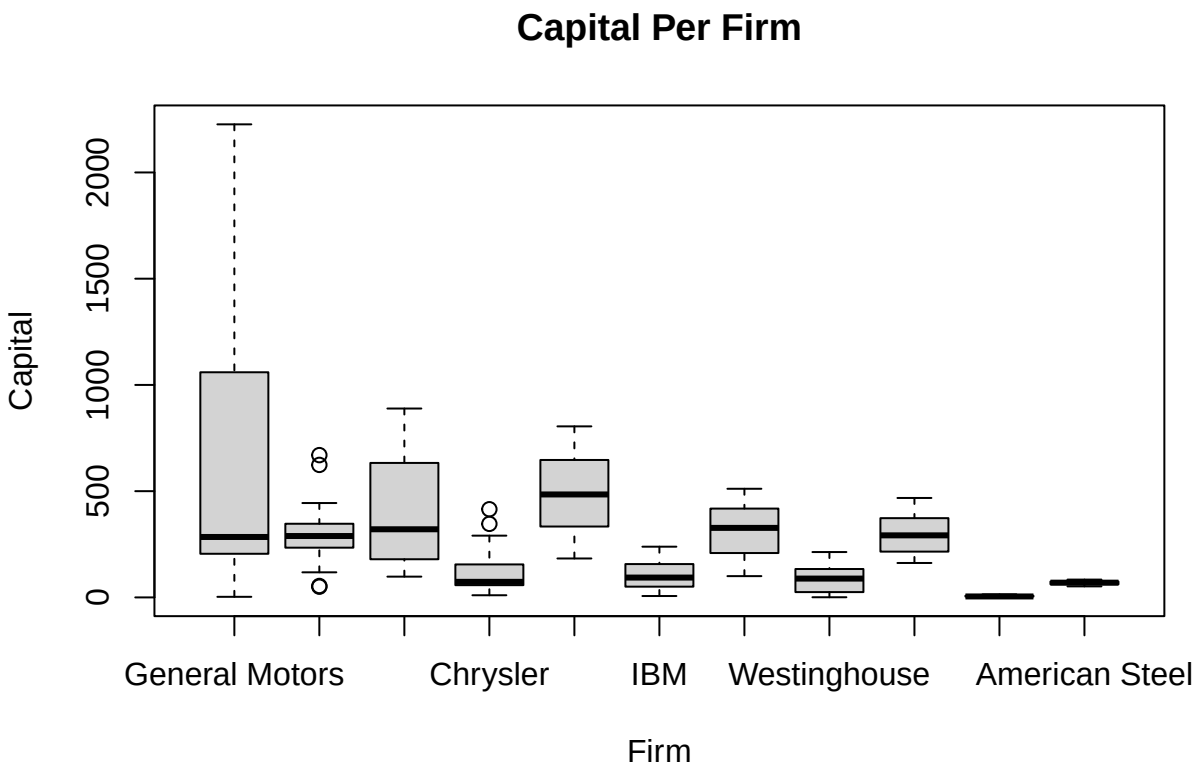
Capital

```
par(cex = 0.5)
par(cex.main = 3)
barplot(height = capital, names = firm, xlab = "Firms", ylab = "Capital",
        main = "Capital Values Per Firm", col = "blue")
```



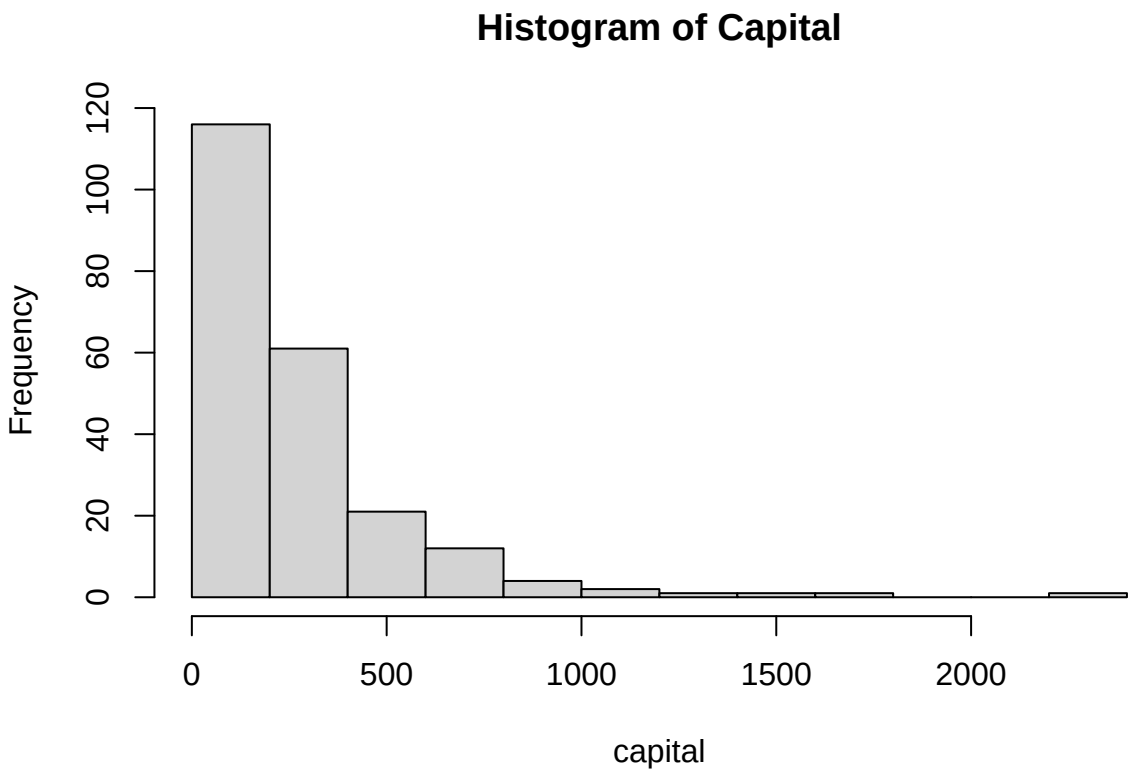
From the barplot, we can see that capital for each firm increases over the years. Most of the firms have roughly comparable magnitudes, except for General Motors, which eclipses all of the other firms about halfway through the panel.

```
plot(firm, capital, main = "Capital Per Firm", xlab = "Firm", ylab = "Capital")
```



Similarly to the boxplot of investment, we can see larger variances associated with large amounts of capital. Most of the firms are roughly comparable, except for the extreme tail on General Motors. There are relatively few outliers in the data.

```
hist(capital, main="Histogram of Capital")
```



Similarly to the investment histogram, the long right tail indicates that most values are clustered towards the lower end of the scale, with a few extreme outliers.

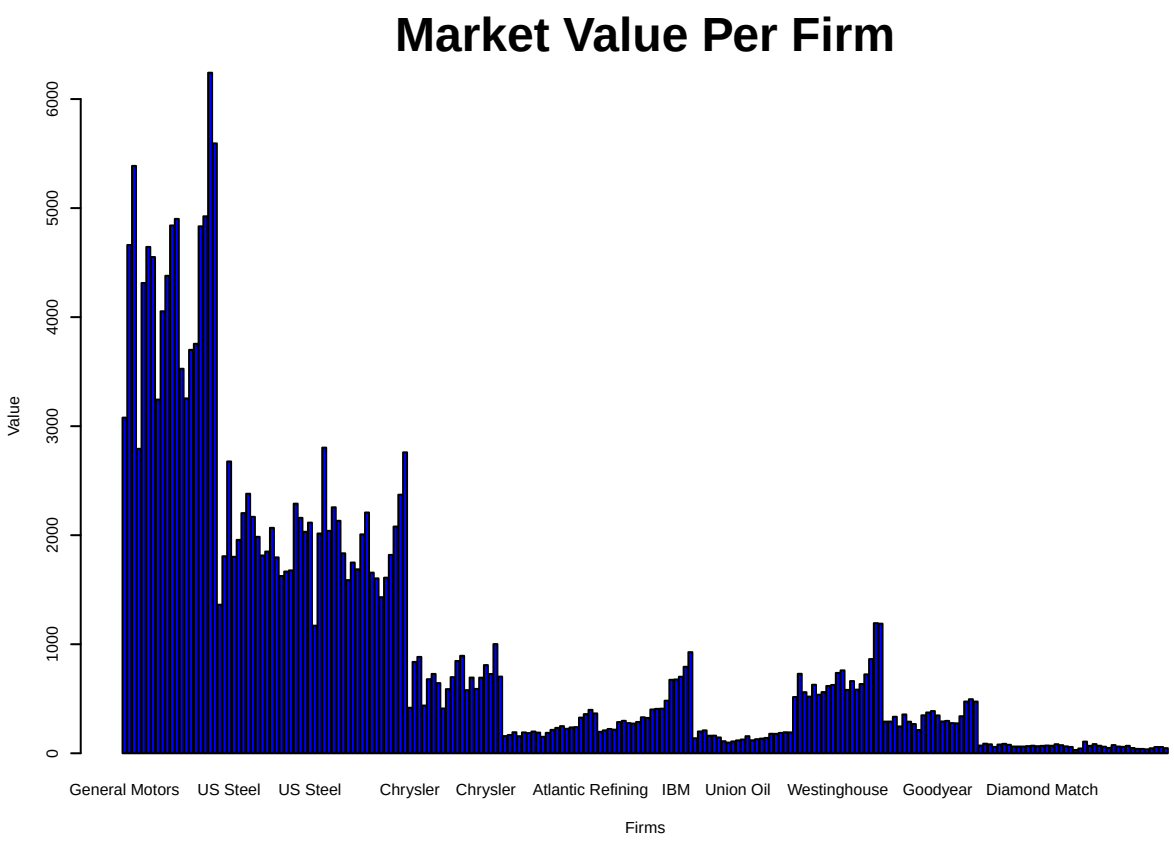
```
fivenum(capital)
```

```
## [1]    0.8   67.1  180.1  345.0 2226.3
```

The five number summary paints a similar picture of capital, with a massive maximum value compared to the IQR. Most of the values in the dataset fall in the low to mid hundreds, with the lowest end being nearly zero. The highest value is over 2,000.

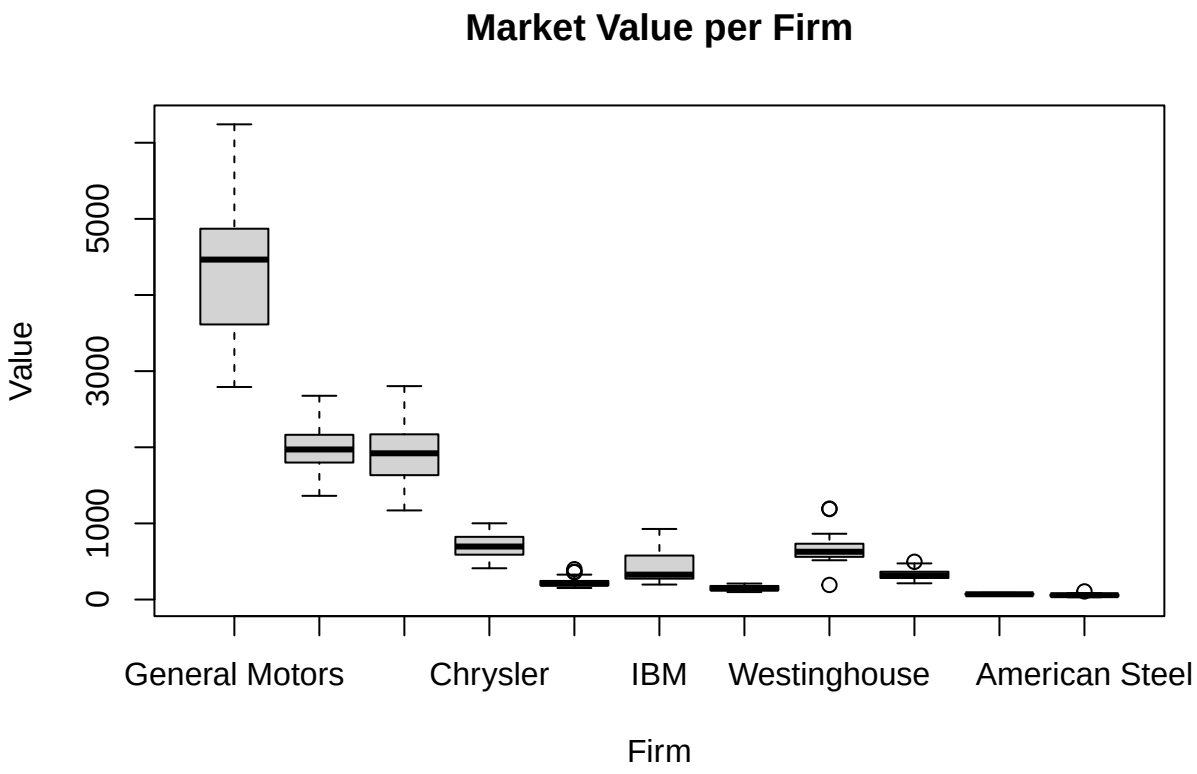
Value

```
par(cex = 0.5)
par(cex.main = 3)
barplot(height = value, names = firm, xlab = "Firms", ylab = "Value", main = "Market Value Per Firm", co
```



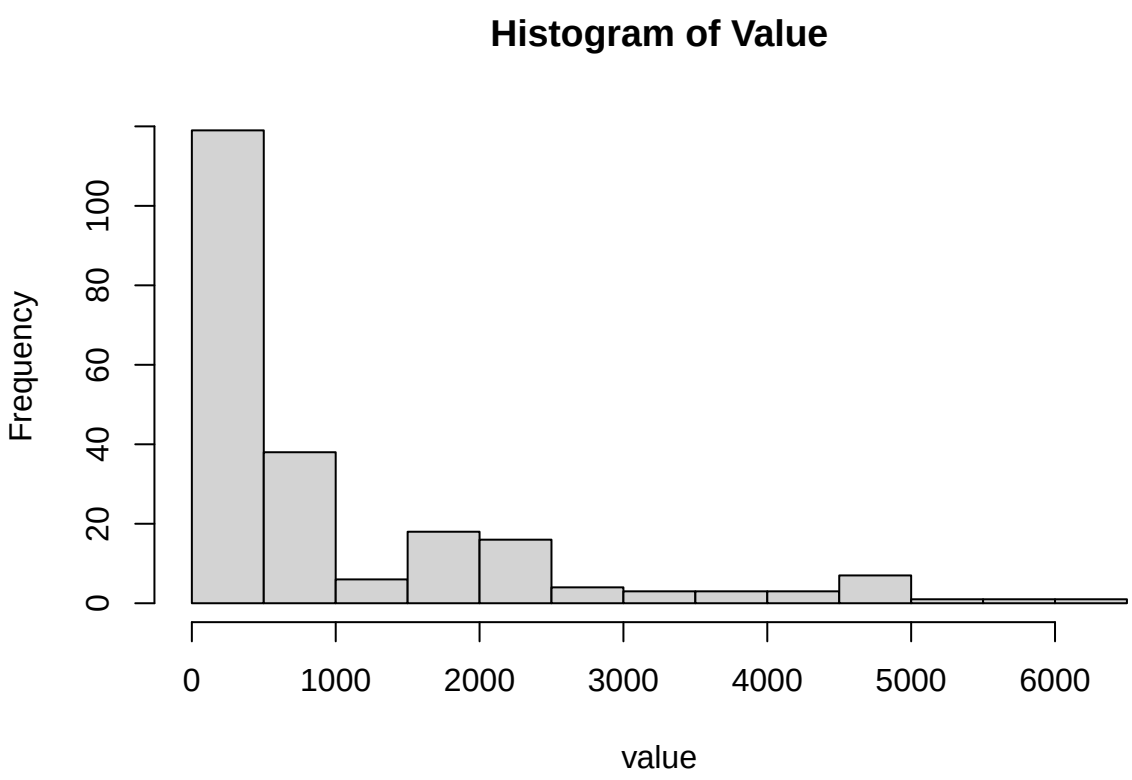
The market value of each firm does increase over time as seen in the barplot, but not as much as in the plots of investment or capital. The largest firm is still General Motors by a significant amount. The lower-valued firms have a stronger, more consistent upward trend, indicating that they are less volatile than the larger firms.

```
plot(firm, value, main = "Market Value per Firm", xlab = "Firm", ylab = "Value")
```



The boxplot shows that General Motors is the largest firm, with accordingly the largest variance in market value.

```
hist(value, main="Histogram of Value")
```



Like with the previous two histograms, the histogram of value shows a long right tail. The tail is more solid than the previous histograms, likely due to the consistently large effect of General Motors.

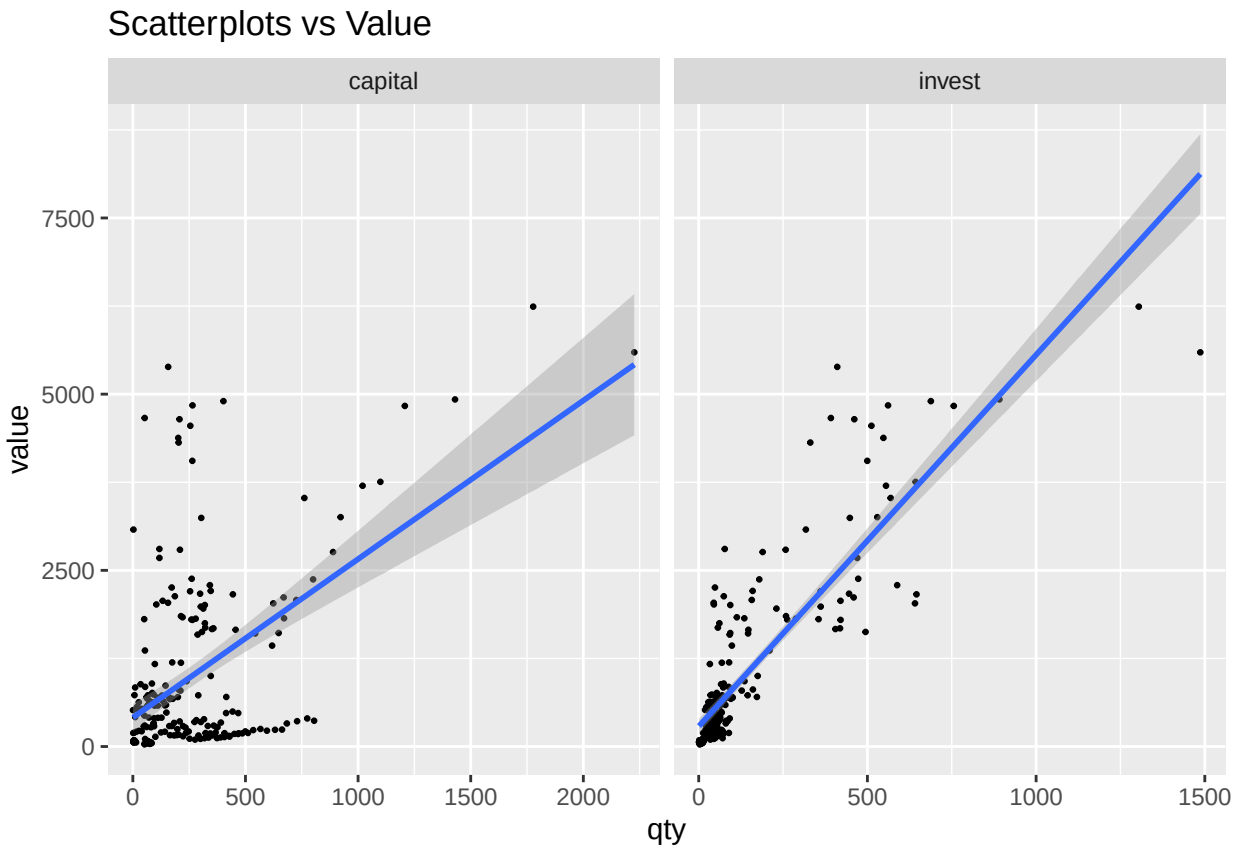
```
fivenum(value)
```

```
## [1]   30.284  159.450  404.650 1607.450 6241.700
```

The five number summary tells the same story as investment and capital, with a long right tail. The IQR for value is much larger than either capital or investment was, indicating a larger spread of the values.

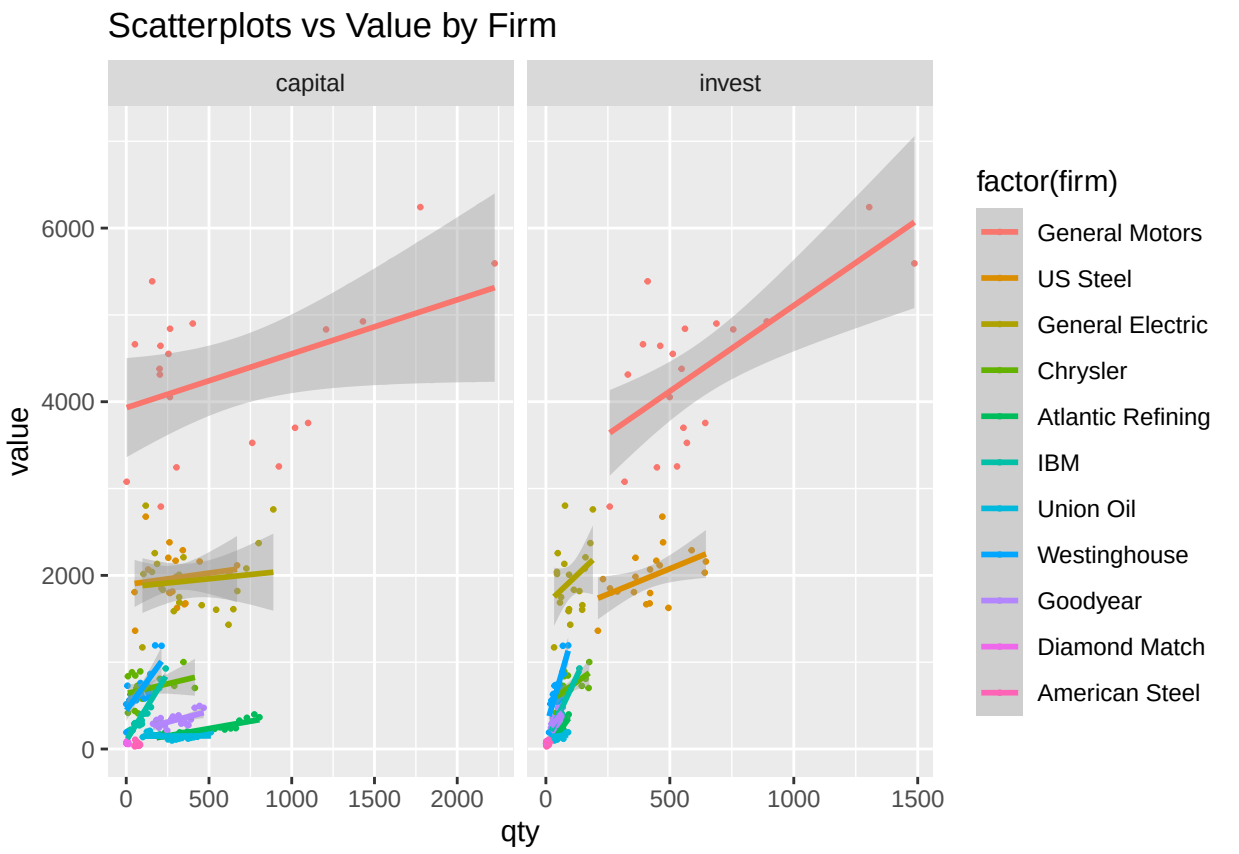
Scatterplots

```
gf_num <- subset(Grunfeld, select=c(invest, value, capital))
gf_num %>%
  tidyr::pivot_longer(c(-value), names_to="key", values_to="qty")  %>%
  ggplot(aes(x=qty, y=value)) +
  geom_point(size=0.5) +
  geom_smooth(formula=y ~ x, method="lm") +
  ggtitle("Scatterplots vs Value") +
  facet_wrap(~ key, scales="free_x")
```



Both scatterplots show a positive association between capital and investment and value. The association in strong for investment than for capital; we can see a cluster of points with low values of capital and high market values that deviate significantly from the rest of the line. The spread of `capital` is larger, especially towards the lower end.

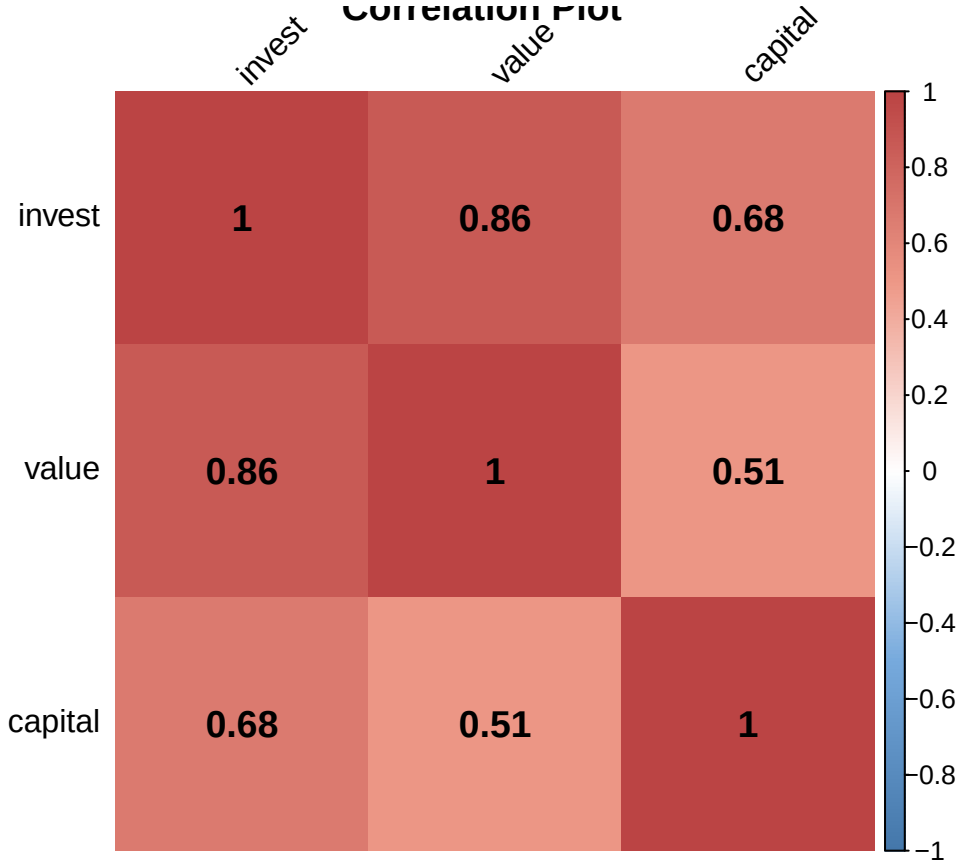
```
cbind(gf_num, firm) %>%
  tidyr::pivot_longer(c(-value, -firm), names_to="key", values_to="qty")  %>%
  ggplot(aes(x=qty, y=value, col=factor(firm))) +
  geom_point(size=0.5) +
  geom_smooth(formula=y ~ x, method="lm") +
  ggtitle("Scatterplots vs Value by Firm") +
  facet_wrap(~ key, scales="free_x")
```



When splitting by firm, we can see that the higher valued firms like GM, GE, and US Steel have much flatter trends than the overall effect observed, and compared to smaller firms. It appears that most of GM's investment trend is dictated by the two data points on the extreme right tail, which drastically change the slope of the line.

Correlation

```
gradient <- colorRampPalette(c("#4477AA", "#77AADD", "#FFFFFF", "#EE9988", "#BB4444"))
corrplot(cor(gf_num), method="shade", shade.col=NA, tl.col="black", tl.srt=45,
         col=gradient(200), addCoef.col="black", title="Correlation Plot",
         number.cex=1.2)
```



We can see large positive correlations between all three of the variables, with the largest being between investment and value and the lowest (but still high) between capital and value. It makes sense that more investment would result in a larger market cap for a firm, as more investment in PP&E is likely to increase the value of the firm to shareholders and result in a substantial return. A larger stock of capital is correlated with a higher value, but the effect is less strong likely because the firm is much more valuable than the sum of its parts, and firms trade more on expectations of future income than on current "book value".

Note that these correlations are aggregated over the entire series of values, across all firms.

3. Modeling

We first convert our dataset into a panel data structure:

```
GrunP <- pdata.frame(Grunfeld, index = c("firm", "year"))
head(GrunP)
```

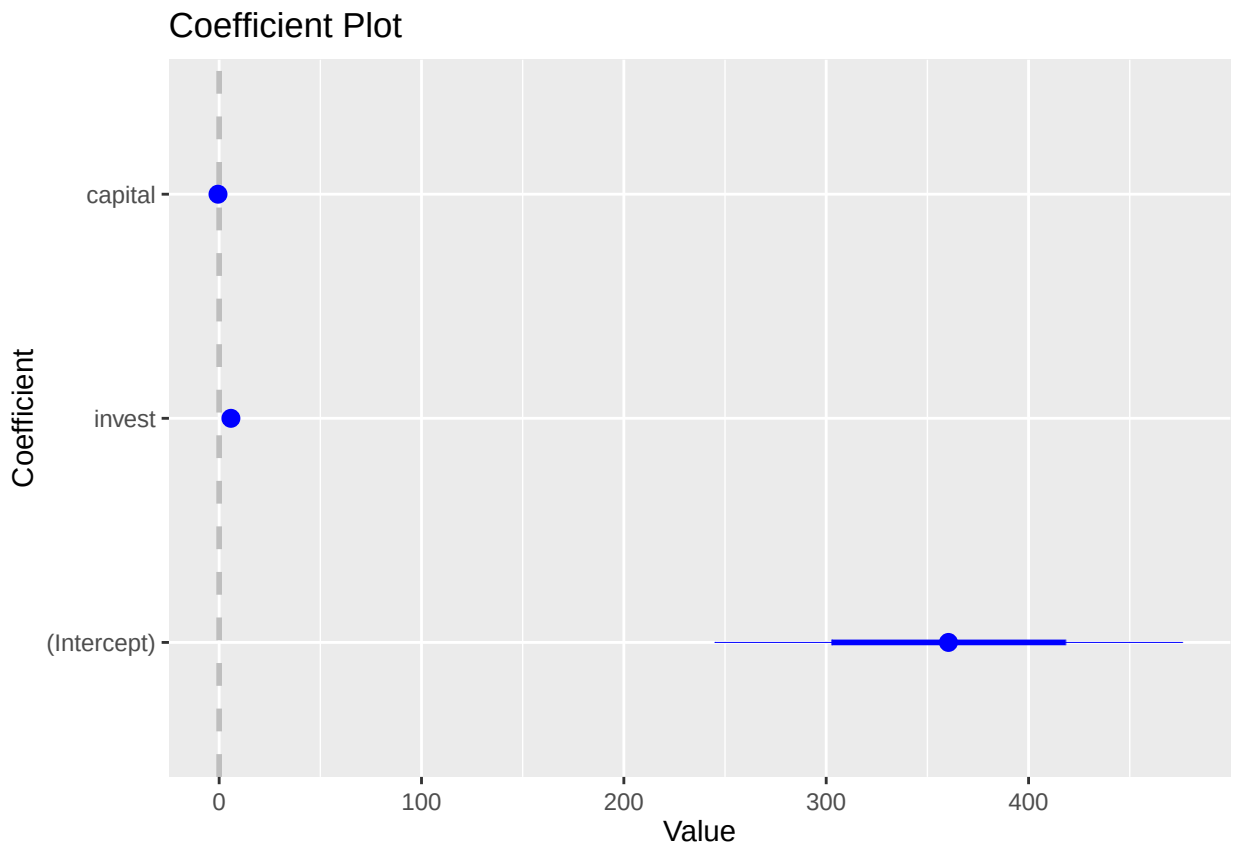
```
##                     invest  value capital           firm year
## General Motors-1935  317.6 3078.5     2.8 General Motors 1935
## General Motors-1936  391.8 4661.7    52.6 General Motors 1936
## General Motors-1937  410.6 5387.1   156.9 General Motors 1937
## General Motors-1938  257.7 2792.2   209.2 General Motors 1938
## General Motors-1939  330.8 4313.2   203.4 General Motors 1939
## General Motors-1940  461.2 4643.9   207.2 General Motors 1940
```

3.1 Pooled Model

```
#Pooled OLS Model
GrunPooled <- plm(value ~ invest + capital, model = "pooling", data = GrunP)
summary(GrunPooled)
```

```
## Pooling Model
##
## Call:
## plm(formula = value ~ invest + capital, data = GrunP, 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
```

```
coefplot(GrunPooled)
```



In the pooled model, both investment and capital are significant at the 5% level. The coefficient of investment is strongly positive, as we expected. Somewhat surprisingly, the coefficient of capital is slightly negative, which indicates that after accounting for the effect of PP$E investment, additional capital has a negative effect on market value.

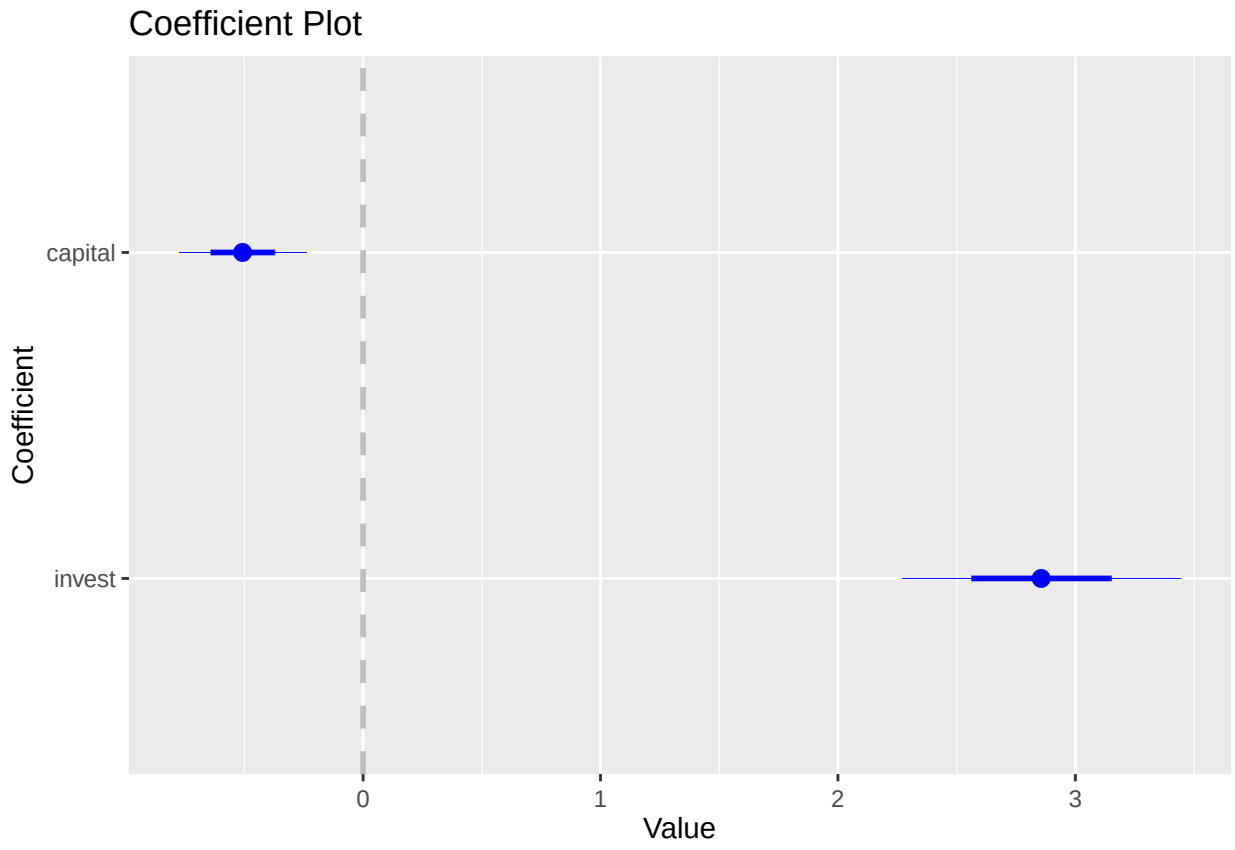
3.2 Fixed Effects

```
#Fixed Effects Model
GrunFixed <- plm(value ~ invest + capital, model = "within", data=GrunP)
summary(GrunFixed)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = value ~ invest + capital, data = GrunP, model = "within")
##
## Balanced Panel: n = 11, T = 20, N = 220
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -807.7945  -73.9885   -5.2818   56.0171 1367.4680
##
## Coefficients:
##         Estimate Std. Error t-value  Pr(>|t|)
```

```
## invest   2.85608    0.29305  9.7461 < 2.2e-16 ***
## capital -0.50787    0.13376 -3.7969 0.0001925 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    23084000
## Residual Sum of Squares: 13582000
## R-Squared:      0.41162
## Adj. R-Squared: 0.37752
## F-statistic: 72.4081 on 2 and 207 DF, p-value: < 2.22e-16
```

```
coefplot(GrunFixed)
```



With the fixed effects model, investment and capital are still significant. However, the effect of `invest` has decreased by nearly half when accounting for the fixed effects. As we saw with the exploratory data analysis before, this is quite possibly due to the large difference in scale between some of the firms.

```
fixef(GrunFixed)
```

```
##    General Motors          US Steel  General Electric          Chrysler
##          2926.609           949.222          1852.405           508.811
## Atlantic Refining               IBM        Union Oil      Westinghouse
##           302.169           314.570           173.804           591.902
##           Goodyear     Diamond Match    American Steel
##           365.305            65.129            72.532
```

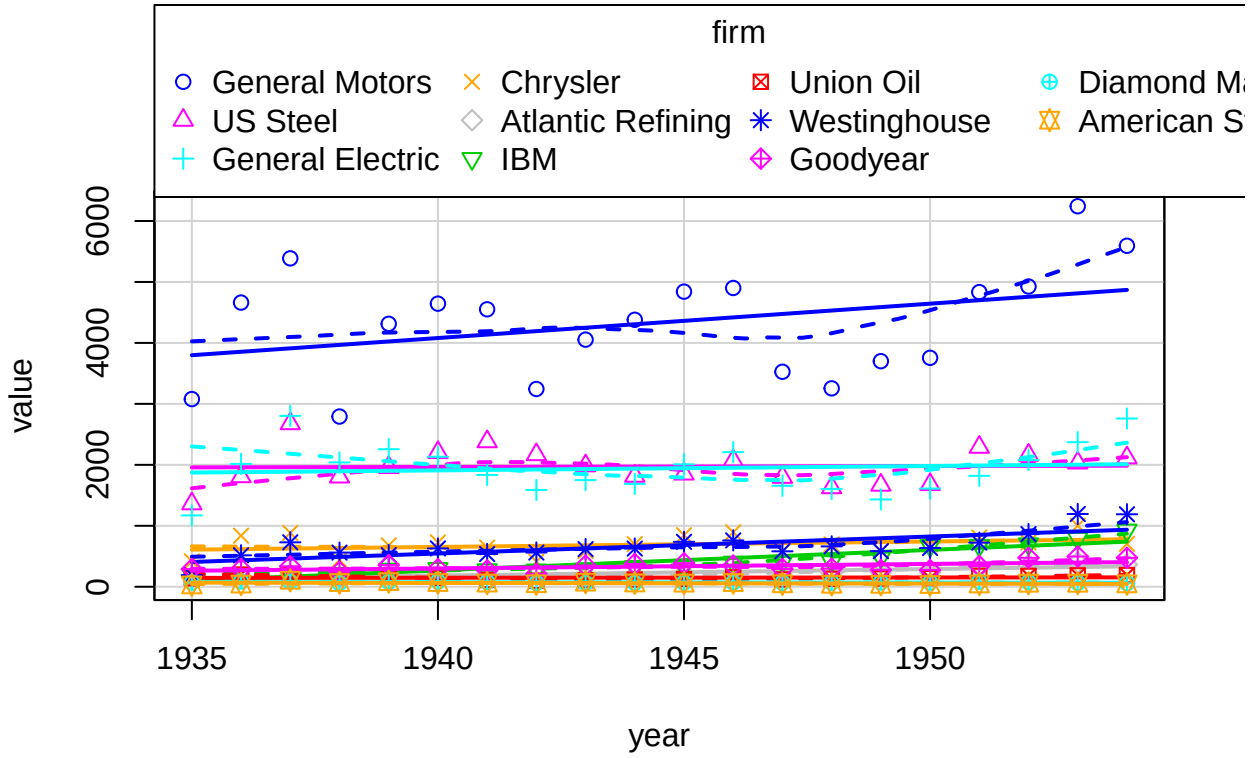
Examining the fixed effects, we can see much larger values for General Motors, US Steel, and General Electric, which we expected based on the data.

```
#Hypothesis Test Between Pooled Effects and Fixed Effects
pFtest(GrunFixed, GrunPooled)
```

```
##
##  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
```

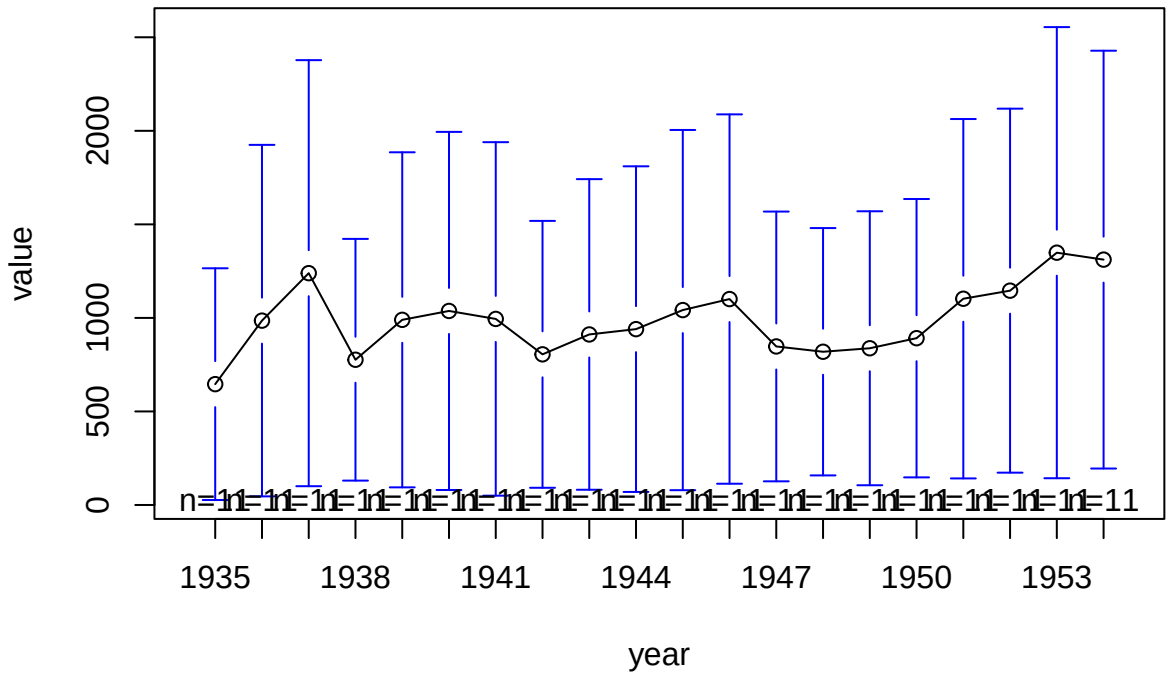
We now perform an F-test to compare the pooled effects and fixed effects estimates. The p-value is effectively zero, indicating that we should reject the null hypothesis in favor of the alternative that the effects are significant.

```
#Heterogeneity Over Time
scatterplot(value ~ year|firm, data = Grunfeld)
```



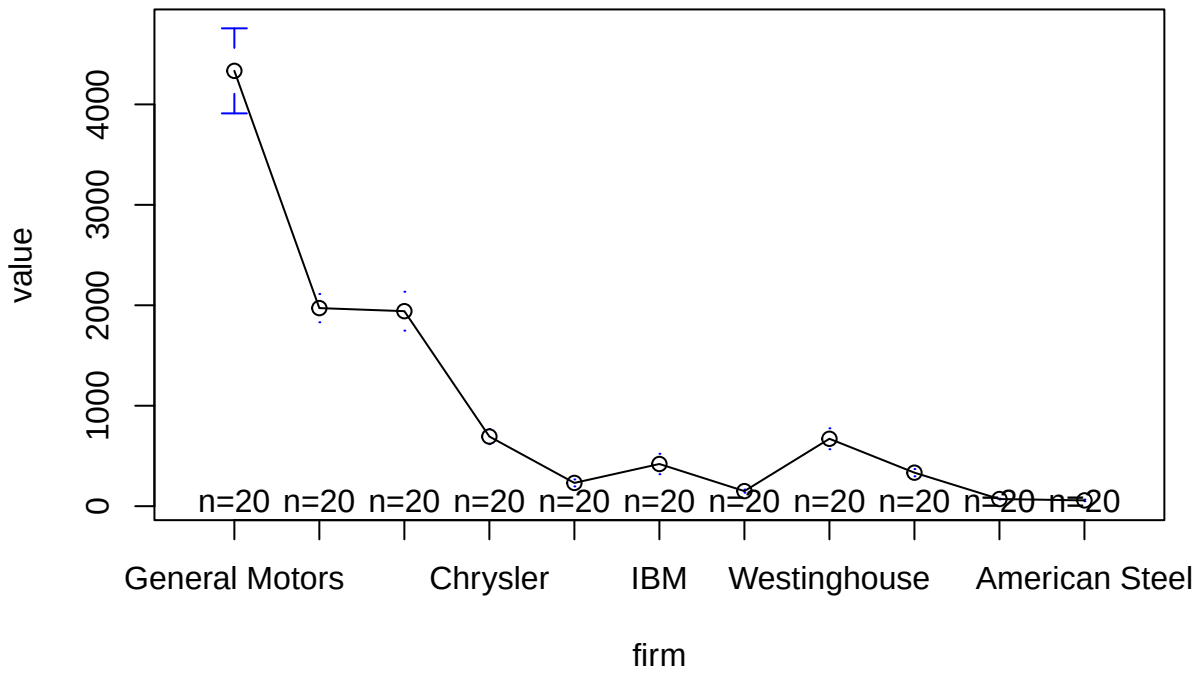
From this plot, we can see that the value over time for most firms is relatively stable. The one exception is GM, which has a more noticeable trend than the other firms. It definitely appears that the firm has a larger effect on value than does time.

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



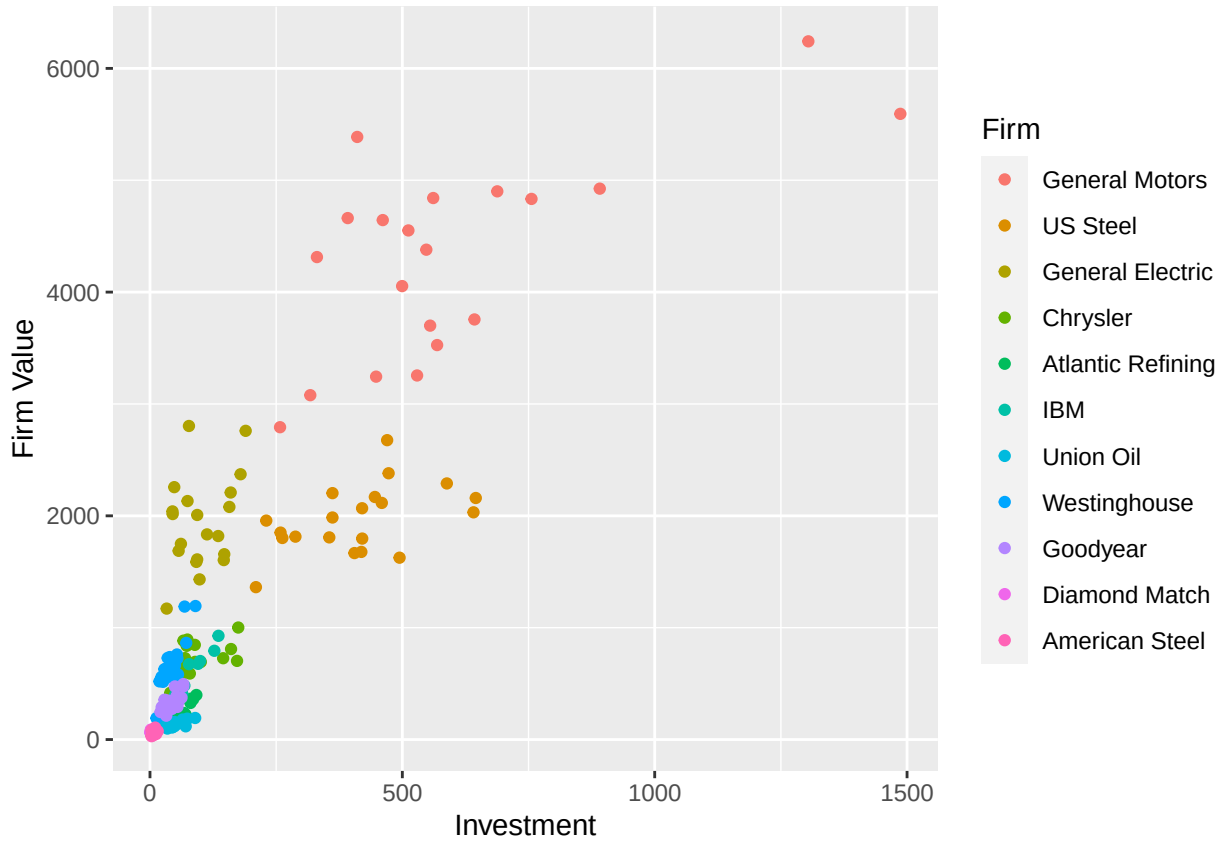
Comparing the means across time for all firms, we can see that the confidence intervals overlap quite significantly even though the mean values are trending up. It appears that the overall mean stay around 1000 across most of the period represented in the dataset.

```
#Heterogeneity Across Firms
plotmeans(value~firm, data=Grunfeld)
```



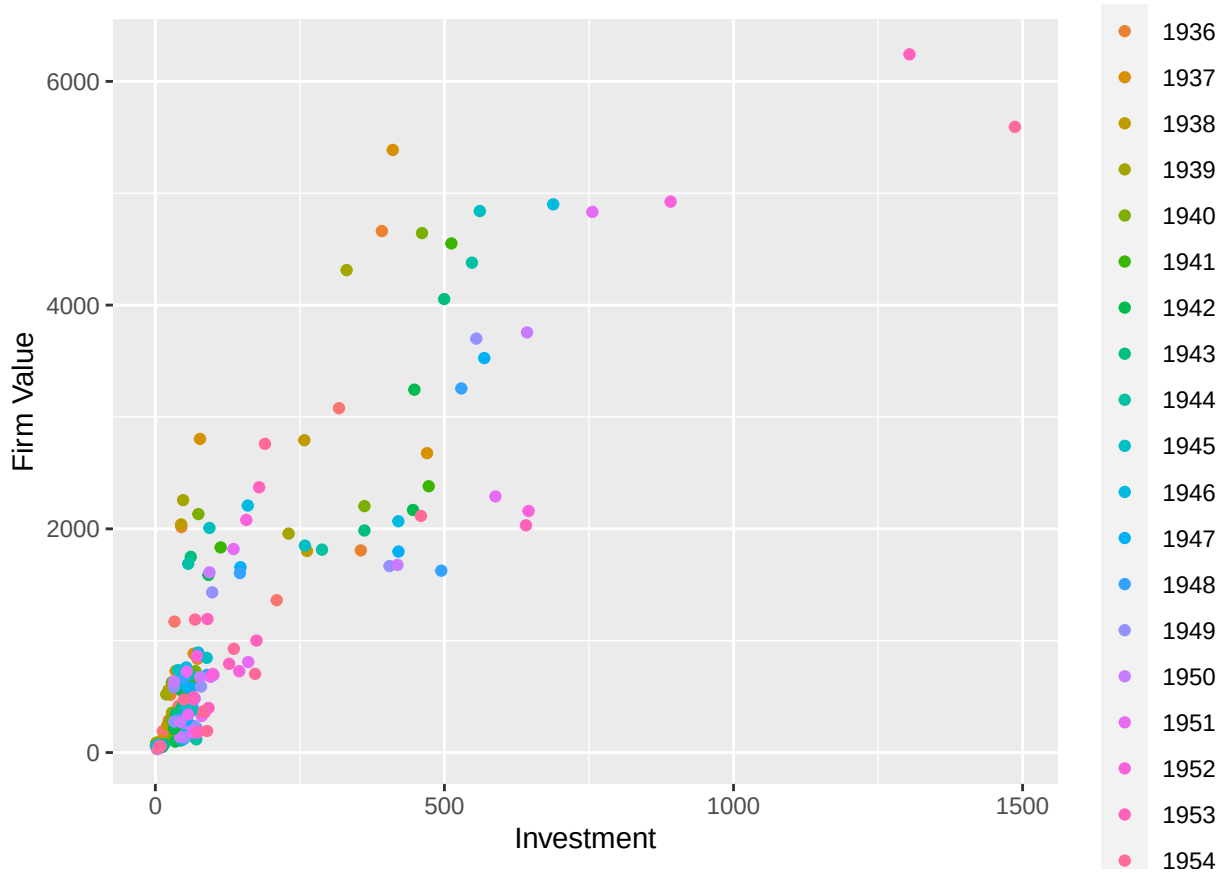
We can see stark differences in the means of each firm, with General Motors far outstripping the rest of the firms.

```
#Visualize value vs. investment by firms
ggplot(Grunfeld, aes(x=invest, y=value, col=factor(firm))) + geom_point() +
  xlab("Investment") + ylab("Firm Value") + scale_color_discrete(name = "Firm")
```



From this plot, it is pretty clear that firms that invest more have a higher firm value. This idea is aligned with the information that we got from the scatterplot, where we saw that GM had noticeable trends. We can see pretty clear clusters representing each firm. Most of the firms are clustered together in the bottom left of the plot, with the larger firms spreading out towards the top.

```
#Visualize value vs. investment by year
ggplot(Grunfeld, aes(x=invest, y=value, col=factor(year))) + geom_point() +
  xlab("Investment") + ylab("Firm Value") + scale_color_discrete(name = "Year")
```



Coloring each point by the year, we can generally see a positive trend as the years increase, although the effect is harder to distinguish towards the bottom of the graph.

All of these plots and our statistical tests point us towards the fixed effects model over the pooled model. We turn now to the random effects model to explore whether it is a more appropriate model than the fixed effects.

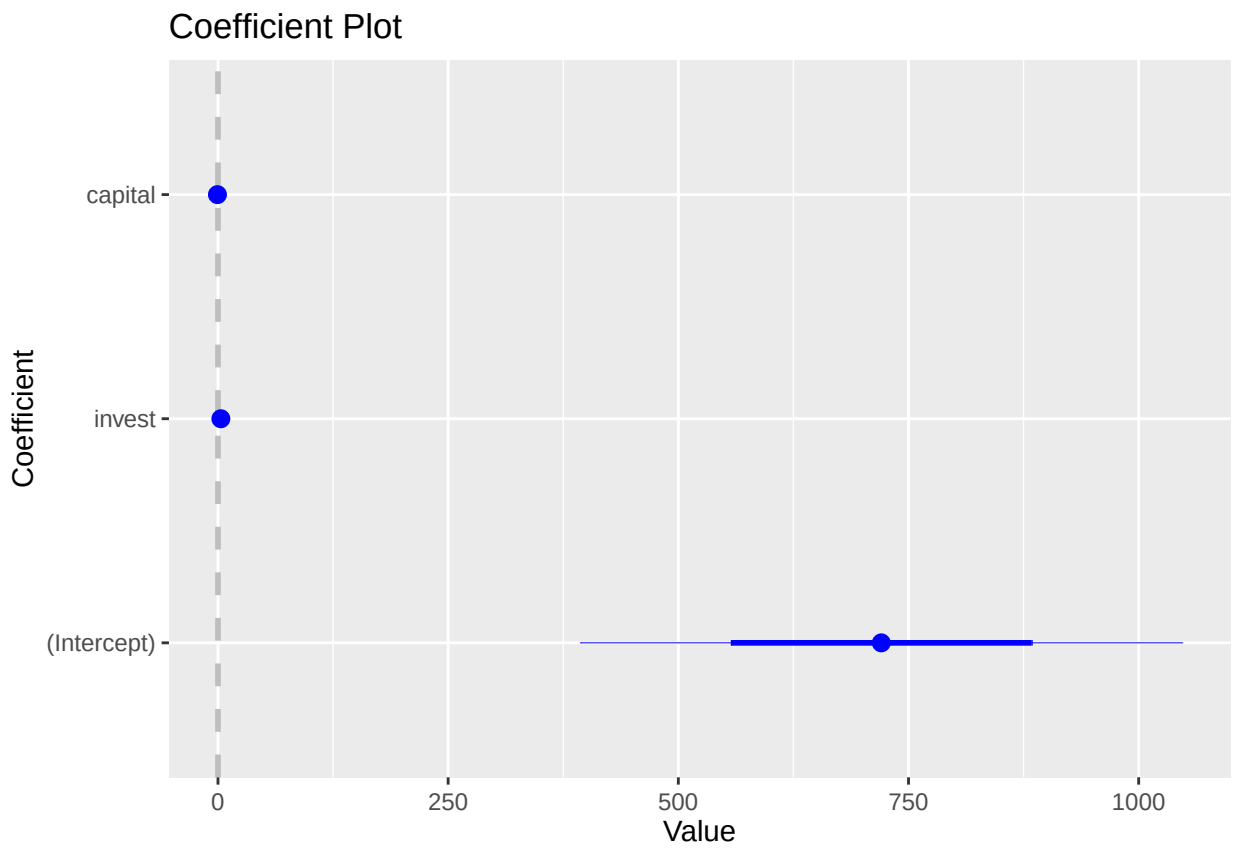
3.3 Random Effects

```
#Random Effects Model
GrunRandom <- plm(value ~ invest + capital, model = "random", data = GrunP)
summary(GrunRandom)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = value ~ invest + capital, data = GrunP, model = "random")
##
## Balanced Panel: n = 11, T = 20, N = 220
##
## Effects:
##                     var  std.dev share
## idiosyncratic  65613.9    256.2 0.201
## individual    261517.9    511.4 0.799
```

```
## theta: 0.8887
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -603.763 -103.038  -59.263   46.043 1617.939
##
## Coefficients:
##              Estimate Std. Error z-value  Pr(>|z|)
## (Intercept) 720.44836  163.37647  4.4097 1.035e-05 ***
## invest        3.13408    0.29423 10.6520 < 2.2e-16 ***
## capital      -0.58217    0.13641 -4.2679 1.973e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    27294000
## Residual Sum of Squares: 15460000
## R-Squared:      0.43359
## Adj. R-Squared: 0.42837
## Chisq: 166.115 on 2 DF, p-value: < 2.22e-16
```

```
#Effects Plot
coefplot(GrunRandom)
```



The random effects model has similar, though slightly larger magnitude, coefficients to the fixed effects model, which is to be expected. Both coefficients are highly significant.

```
#Hypothesis test between Random and Fixed Effects
phtest(GrunFixed, GrunRandom)
```

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

Performing a Hasuman test, the p-value approximately equal to zero indicates that we should reject the null hypothesis in favor of the alternative hypothesis that the fixed effects model is inconsistent and the random effects model is better. This is not surprising, as we can imagine a number of reasons that random effects would be present in the system. Market value is a noisy estimator of the true underlying value of a firm, so it makes sense that the intercept terms would be random.

Based on the results of our statistical hypothesis tests, we conclude that the best model for this panel data is the Random Effects model.

Overall, through our analysis we can conclude that there is a relationship between investment, capital, and the value of a firm. We were able to demonstrate, quite logically, that a firm that invests more increases their value. We visualized the effect of each firm and each year and discovered that fixed effects were more appropriate than a pooled model. Larger firms, like GM, also had larger investments. Notably, we found that larger values of capital stock (holding investment constant) predicted a lower market value, which warrants future study. We finally concluded, based on statistical testing, that the random effects model is the most appropriate model for this panel dataset.

Qualitative Dependent Variable

1. Dataset

For this section, we will be examining NBA player statistics from the 1997-98 season up through the 2022-23 season (as of March 8, 2023). The statistics are sourced from basketball-reference.com.

The question we wish to examine is which player statistics are most predictive of a player changing teams (either voluntarily or through being traded). NBA teams are extremely valuable businesses, and their decisions can have large economic impacts. Understanding the dynamics of team changes can give us insight into what makes a valuable NBA player and a successful team.

```
nba_df <- read.csv("NBA_Player_Stats.csv")
nba_df[is.na(nba_df)] <- 0.0
```

All of the null values in this dataset occur for scoring percentages, where the divisor (shots attempted) is zero. For these values, it is reasonable to impute 0.

```
nba_df$Traded <- 0
nba_df_traded <- nba_df[nba_df$Tm == "TOT",]
# Construct our binary dependent variable
for (i in 1:dim(nba_df_traded)[1]) {
  player <- nba_df_traded[i,"Player"]
  season <- strsplit(nba_df_traded[i,"Season"], "-")[[1]]
  season_yr <- as.numeric(season[1])
  for (lag in 0:2) {
    prev_season <- paste(season_yr - lag - 1,
                         substr(as.character(season_yr - lag), 3, 4), sep="-")
    # Set traded for the previous year
    nba_df[nba_df$Season == prev_season & nba_df$Player == player, "Traded"] <- 1
  }
}
# Drop partial rows for traded players
for (i in 1:dim(nba_df_traded)[1]) {
  player <- nba_df_traded[i,"Player"]
  season <- nba_df_traded[i,"Season"]
  nba_df <- nba_df[nba_df$Season != season | nba_df$Player != player | nba_df$Tm == "TOT",]
}
```

We construct our binary dependent variable `Traded`, with a 1 indicating that a player changes teams within the next 3 seasons. Concretely, the question we are seeking to answer is: "Given an NBA player's season statistics, how likely is it that he will change teams within the next 3 seasons?"

```
nba_df <- subset(nba_df, select=c("Age", "G", "FGPct", "X3PPct", "eFGPct",
                                  "FTPct", "TRB", "AST", "STL", "PTS", "Traded"))
head(nba_df)
```

```
##   Age  G FGPct X3PPct eFGPct FTPct TRB AST STL  PTS Traded
## 1  28 31 0.377  0.161  0.386 1.000 1.2 1.9 0.5  7.3      0
## 2  23 59 0.403  0.211  0.409 0.672 2.0 0.9 0.6  6.4      1
## 3  21 82 0.485  0.412  0.493 0.784 7.1 2.6 1.1 22.3      0
## 4  24 60 0.428  0.375  0.510 0.784 2.4 3.5 1.2  8.1      0
## 7  22 82 0.428  0.364  0.479 0.875 4.9 4.3 1.4 19.5      0
## 8  23 66 0.408  0.202  0.422 0.873 2.8 3.4 1.3 11.7      0
```

For this project, we will use a subset of 10 predictors:

- `Age`: The player's age, in years
- `G`: Number of games played in this season
- `FGPct`: Percentage of field goals (2-point baskets) scored vs attempted
- `X3PPct`: Percentage of 3-point field goals scored vs attempted
- `eFGPct`: Effective field goal percent (after adjusting for the fact that 3-pointers are worth more than 2-pointers)
- `TRB`: Total rebounds per game
- `AST`: Assists per game
- `STL`: Steals per game
- `PTS`: Average points scored per game

At this time, we will also remove some outliers with shooting percentages of 0 or 1 (due to an inadequate number of games played).

```
nba_df <- nba_df[nba_df$FGPct > 0.0 & nba_df$FGPct < 1.0,]
```

2. Descriptive Analysis

Statistical Summary

```
summary_df <- sapply(nba_df, fivenum)
rownames(summary_df) <- c("Min", "Q1", "Median", "Q3", "Max")
summary_df <- t(summary_df)
stargazer(summary_df, type="text")
```

Five Number Summary

```
##
## ======================================
##          Min   Q1   Median   Q3    Max
## --------------------------------------
## Age      18    23     26     30     44
## G         1    33     57     73     85
## FGPct  0.071 0.400 0.439  0.481  0.857
## X3PPct   0     0   0.311  0.368    1
## eFGPct 0.071 0.446 0.488  0.528    1
## FTPct    0   0.654 0.750  0.818    1
## TRB      0   1.800   3    4.800  16.300
## AST      0   0.600 1.200  2.500  11.700
## STL      0   0.300 0.600  0.900  2.900
## PTS    0.100 3.800 6.900  11.600 36.100
## Traded   0     0     0      0      1
## --------------------------------------
```

The five number summary shows that most players are in their 20s, with a couple falling outside of this range. Most players played several dozen games, and scored fewer than half of their attempted shots. Rebounds, assists, steals, and points all have seemingly large max values compared to their interquartile ranges, which may be indicative of outliers. Finally, most players do not change teams frequently.

```
stargazer(nba_df, type="text")
```

Mean and Std. Dev

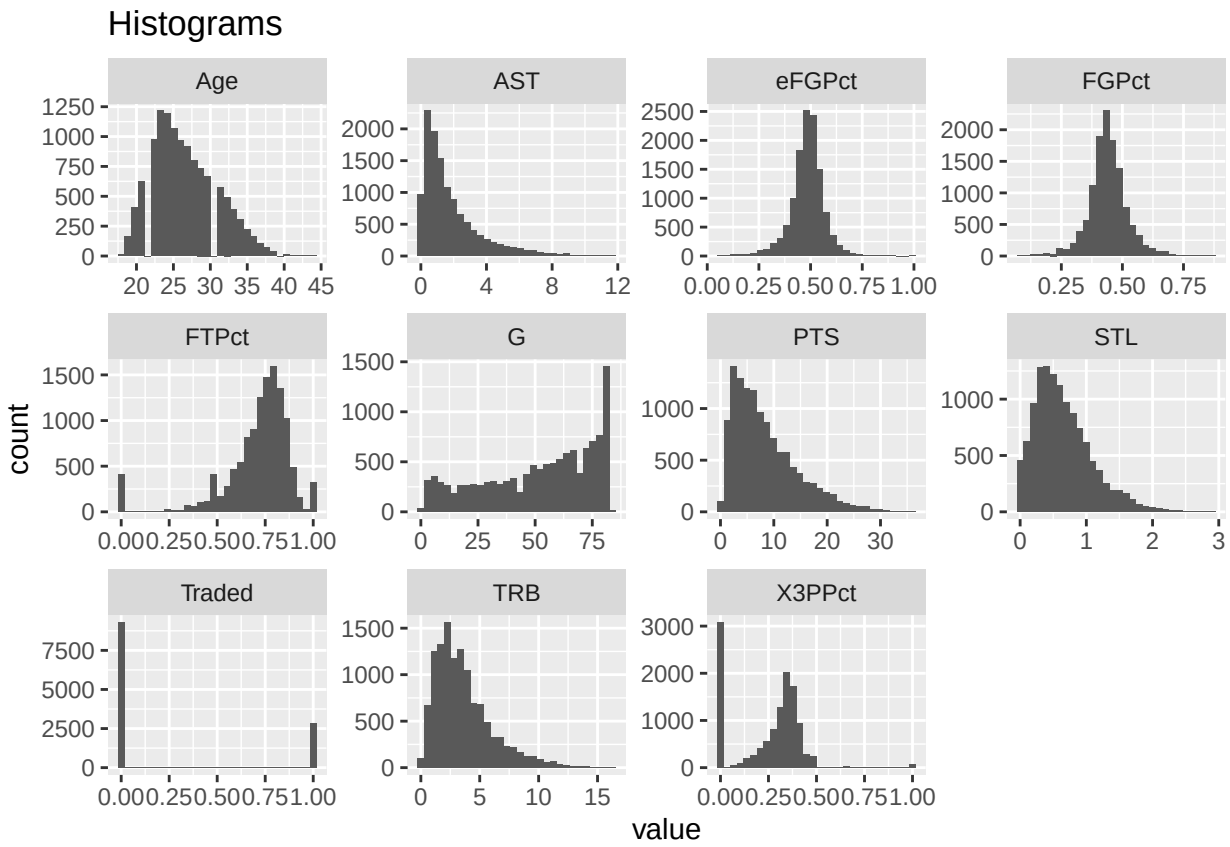
```
##
## ===============================================
## Statistic   N     Mean  St. Dev.  Min   Max
## -----------------------------------------------
## Age       12,139 26.587  4.341    18     44
## G         12,139 51.692  24.240    1     85
## FGPct     12,139 0.441   0.080   0.071 0.857
## X3PPct    12,139 0.251   0.172   0.000 1.000
## eFGPct    12,139 0.483   0.080   0.071 1.000
## FTPct     12,139 0.709   0.184   0.000 1.000
## TRB       12,139 3.615   2.460   0.000 16.300
## AST       12,139 1.852   1.800   0.000 11.700
## STL       12,139 0.648   0.435   0.000 2.900
## PTS       12,139 8.363   5.982   0.100 36.100
## Traded    12,139 0.235   0.424     0     1
## -----------------------------------------------
```

The mean and standard deviation confirm our observations, especially with regard to scoring percentages. The large standard deviations compared to the mean values of `TRB`, `AST`, `STL`, and `PTS` indicate a wide spread in the data. The mean value of `Traded` tells us that about 23.5% of players change teams in a three year period. There is a large class imbalance in our data, but it may still be possible to extract some useful information from our models.

Distributions

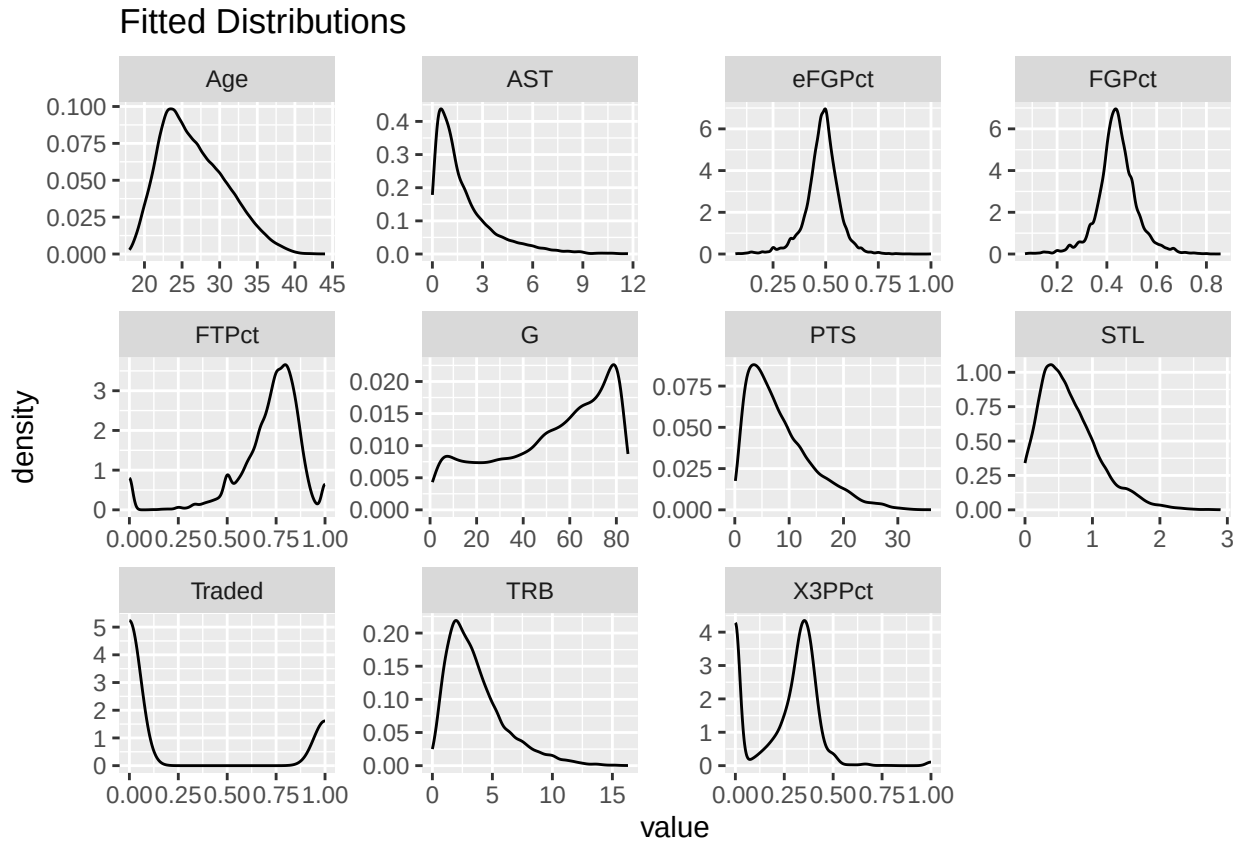
```
nba_df %>%
  tidyr::pivot_longer(everything(), names_to="key")  %>%
  ggplot(aes(x=value)) +
  geom_histogram() +
  ggtitle("Histograms") +
  facet_wrap(~ key, scales="free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plots of histograms confirm the long right tails we suspected on `TRB`, `AST`, `STL`, and `PTS`, which look closer to log-normal or log-logistically distributed. Surprisingly, the values for the shooting percentages look relatively close to normally distributed. Quite a lot of player (over 3000) never attempted or scored a three-point shot.

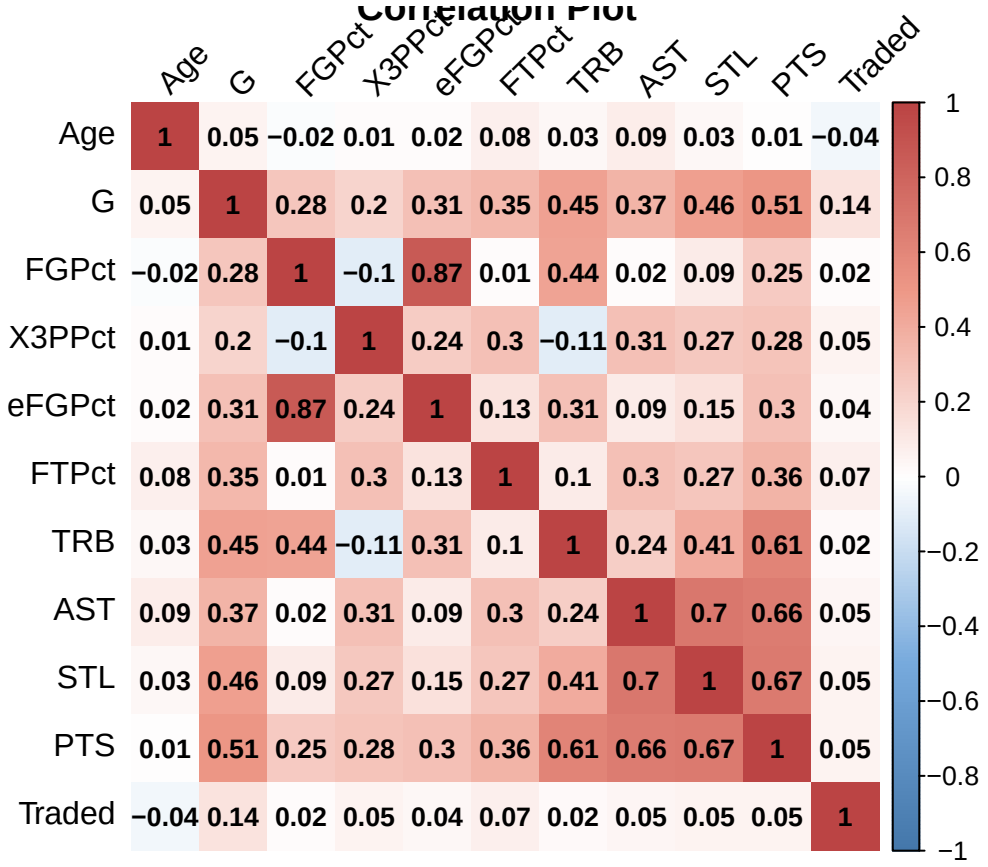
```
nba_df %>%
  tidyr::pivot_longer(everything(), names_to="key")  %>%
  ggplot(aes(x=value)) +
  geom_density() +
  ggtitle("Fitted Distributions") +
  facet_wrap(~ key, scales="free")
```



The fitted distributions confirm our observations on the histograms. `Age` peaks in the early 20s. Number of games played rises, so it appears that many more players played most games in the season than the number who player fewer than half.

Correlation

```
gradient <- colorRampPalette(c("#4477AA", "#77AADD", "#FFFFFF", "#EE9988", "#BB4444"))
corrplot(cor(nba_df), method="shade", shade.col=NA, tl.col="black", tl.srt=45,
         col=gradient(200), addCoef.col="black", title="Correlation Plot",
         number.cex=0.8)
```

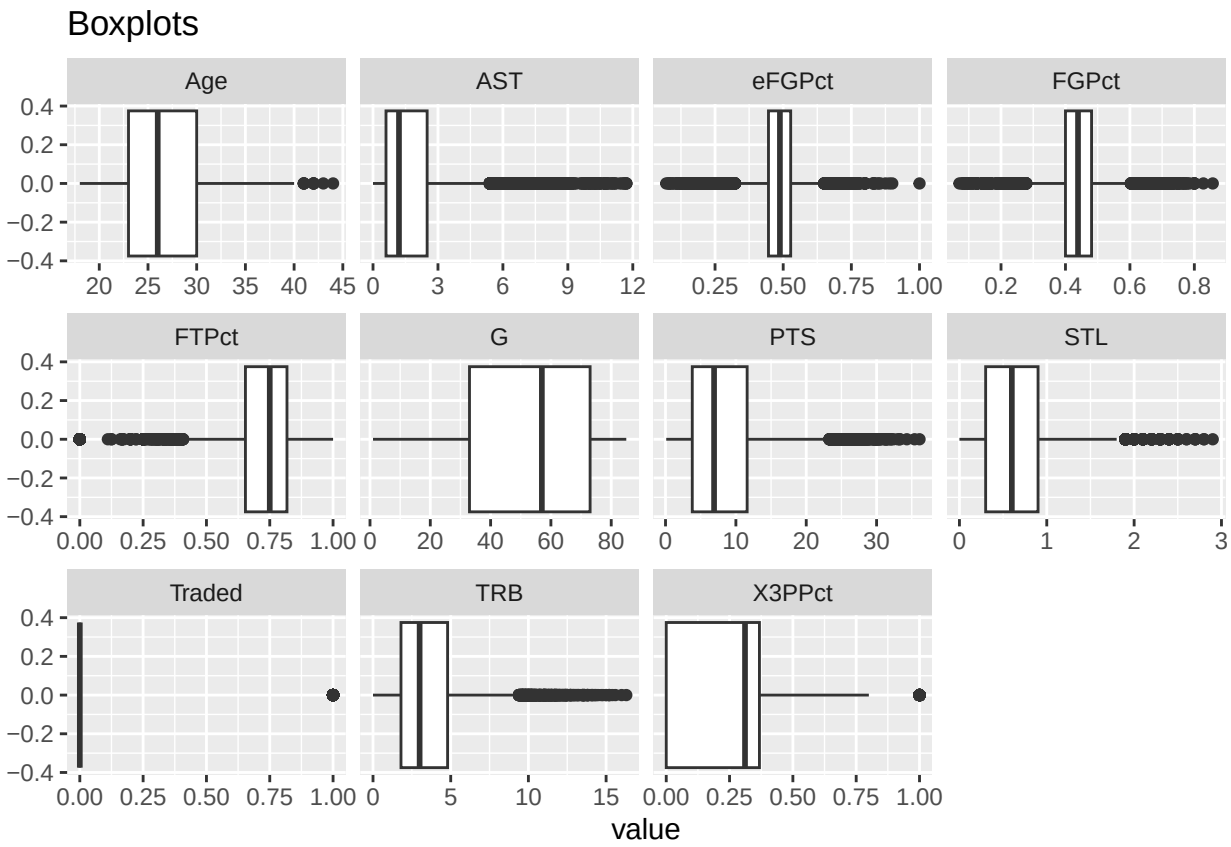


Most of the variables in the dataset have low to moderate correlations. The largest cluster is found between assists per game, steals per game, and points scored per game. `eFGPct` and `FGPct` are highly correlated (unsurprisingly), which raises the specter of colinearity. For best interpretability, we should remove one of the two columns. Also unsurprisingly, games played is positively correlated with all of the other variables, indicating that better players are likely to play in more games.

The correlation between `Traded` and all of the predictors is relatively low, with the highest being 0.14. This indicates that it will be difficult to obtain good predictions using these predictors. This is not surprising, as predicting NBA trades is very valuable. We intentionally chose this because it was a difficult prediction problem, so we are not expecting to achieve stellar results.

Boxplots

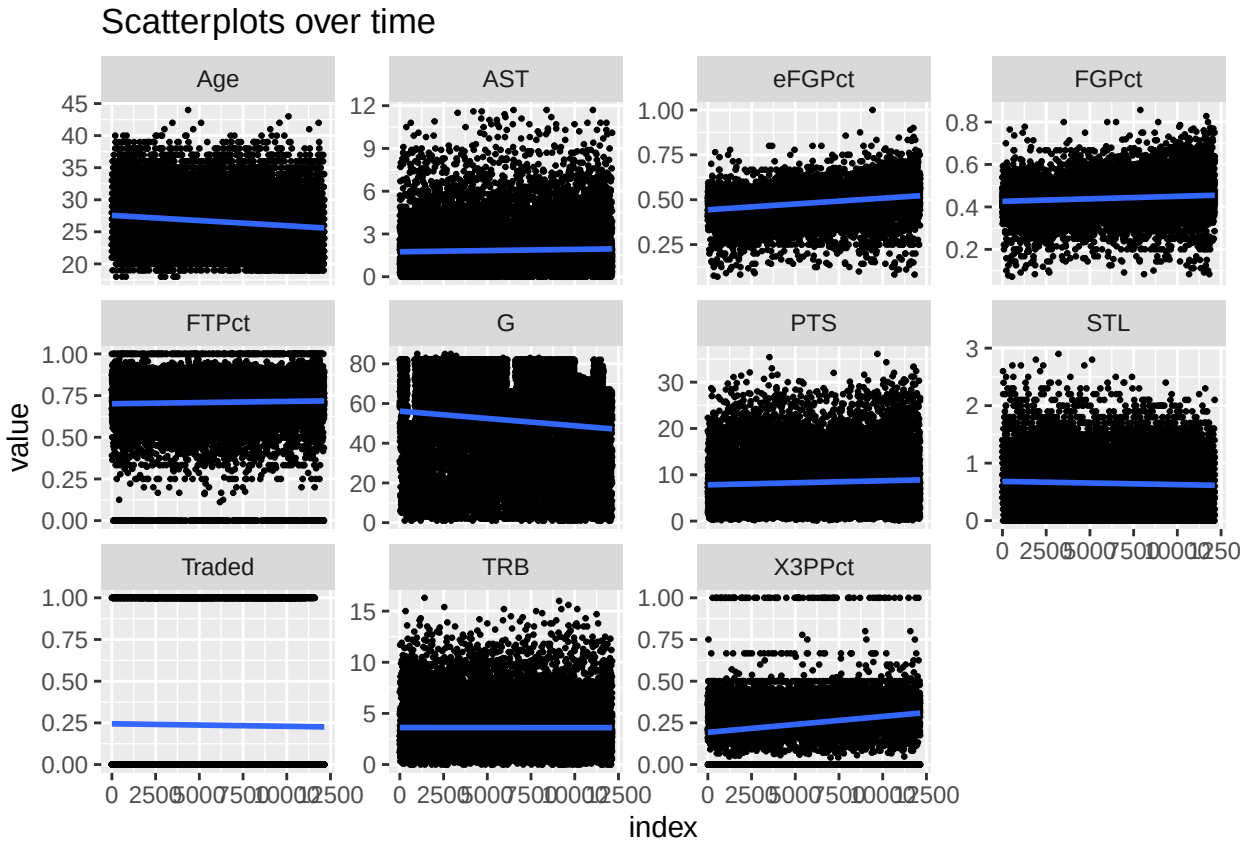
```
nba_df %>%
  tidyr::pivot_longer(everything(), names_to="key")  %>%
  ggplot(aes(x=value)) +
  geom_boxplot() +
  ggtitle("Boxplots") +
  facet_wrap(~ key, scales="free_x")
```



The boxplots indicate long right tails on the four variables previously observed. We can also see that `FGPct` and `eFGPct` have some outliers at the extreme ends, indicating potentially heavier tails than a normal distribution. The longest right tail for points is much, much greater than the median value; the best players score more than three times the points per game that the median player does.

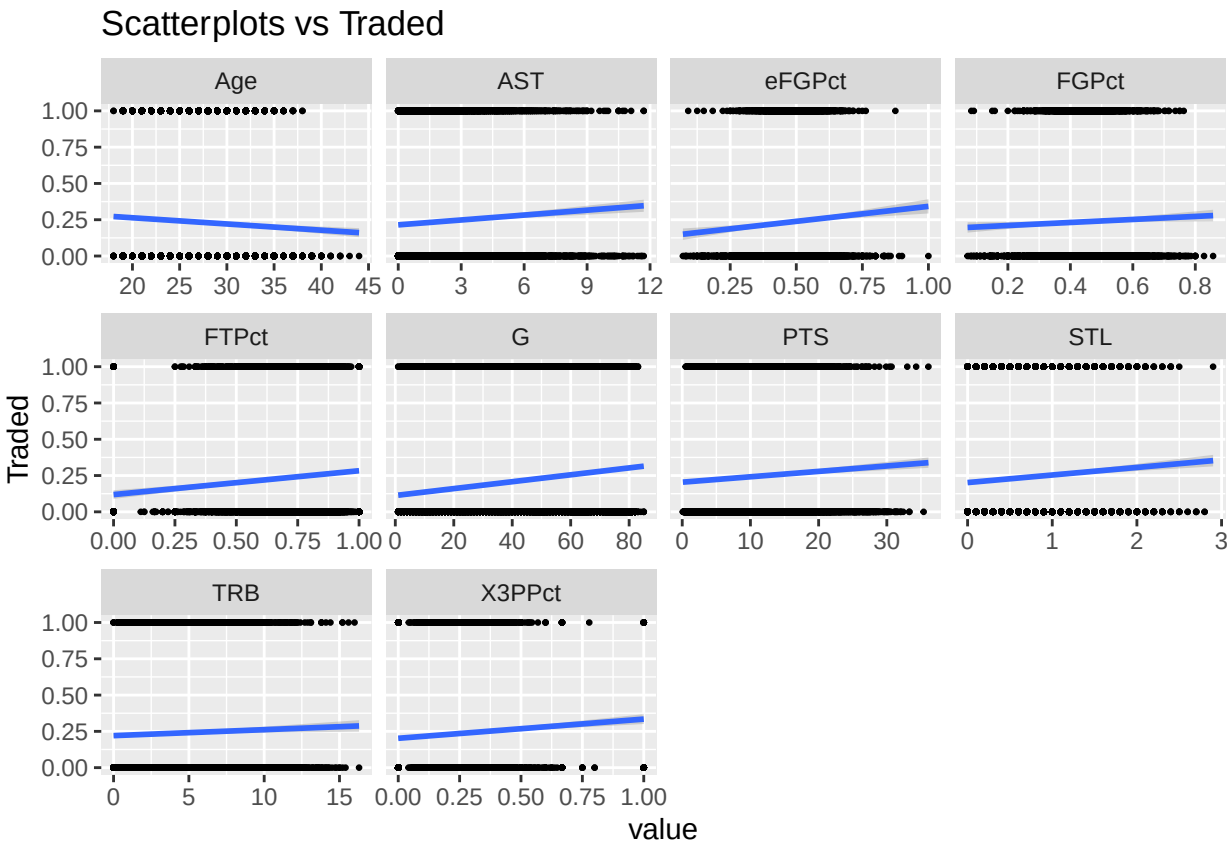
Scatterplots

```
nba.dfi <- cbind(nba_df, 1:12139)
colnames(nba.dfi) <- append(colnames(nba_df), "index")
nba.dfi %>%
  tidyr::pivot_longer(c(-index), names_to="key")  %>%
  ggplot(aes(y=value, x=index)) +
  geom_point(size=0.5) +
  geom_smooth(formula=y ~ x, method="lm") +
  ggtitle("Scatterplots over time") +
  facet_wrap(~ key, scales="free_y")
```



From the scatterplots over time (using index as a proxy), we can make an interesting observation: `eFGPct` is rising over time, faster than `FGPct`. This is likely due to the increase in `X3PPct` over time. Since the late 1990s, NBA players have gotten more accurate with their 3-point baskets, and have made more of them, driving up their effective scoring percentages. Most of the other predictors are relatively flat over time.

```
nba_df %>%
  tidyr::pivot_longer(c(-Traded), names_to="key")  %>%
  ggplot(aes(x=value, y=Traded)) +
  geom_point(size=0.5) +
  geom_smooth(formula=y ~ x, method="lm") +
  ggtitle("Scatterplots vs Traded") +
  facet_wrap(~ key, scales="free_x")
```



Plotting each of the predictors against `Traded`, we can see quite significant overlaps in the data. It appears like there is no single variable that is a good standalone predictor of whether a player will be traded or not. Of particular note is that no players were traded after about age 35; this is around the time when most players start to retire, so it is likely that the only players still playing after this time are exceptional superstars who are very unlikely to be traded.

3. Modeling

3.1 Linear Probability Model

```
# First estimate using OLS to find an estimate of sigma^2
linear.OLS <- lm(Traded ~ ., data=nba_df)
# Re-estimate using feasible GLS
linear.mod <- lm(Traded ~ ., weights=1/resid(linear.OLS)^2, data=nba_df)
summary(linear.mod)
```

```
##
## Call:
## lm(formula = Traded ~ ., data = nba_df, weights = 1/resid(linear.OLS)^2)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25223 -0.17709 -0.13505  0.02219  1.48459
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.009e-02  8.227e-03   2.442  0.01462 *
## Age         -5.743e-04  1.976e-04  -2.906  0.00367 **
## G            8.044e-04  5.705e-05  14.099  < 2e-16 ***
## FGPct       -1.081e-02  4.212e-02  -0.257  0.79747
## X3PPct       1.508e-02  7.742e-03   1.948  0.05141 .
## eFGPct       4.884e-03  4.153e-02   0.118  0.90640
## FTPct       -8.753e-03  3.880e-03  -2.256  0.02410 *
## TRB         -1.828e-03  5.956e-04  -3.070  0.00215 **
## AST          2.479e-04  1.201e-03   0.206  0.83649
## STL         -1.116e-03  3.792e-03  -0.294  0.76860
## PTS          2.677e-04  3.710e-04   0.722  0.47057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6408 on 12128 degrees of freedom
## Multiple R-squared:  0.03896,    Adjusted R-squared:  0.03816
## F-statistic: 49.16 on 10 and 12128 DF,  p-value: < 2.2e-16
```

From the linear probability model, we can see that `Age`, `G`, `FTPct`, and `TRB` are significant at the 5% level. Surprisingly, neither `eFGPct` nor `FGPct` is significant, and neither is `PTS`. We would expect those to be the most predictive of a player being traded to another team. Most of the coefficients have a small magnitude, indicating that the overall effect is weak (even if significant).

3.2 Probit Model

```
probit.mod <- glm(Traded ~ ., family=binomial(link="probit"), data=nba_df)
summary(probit.mod)
```

```
##
## Call:
## glm(formula = Traded ~ ., family = binomial(link = "probit"),
##     data = nba_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0276  -0.7895  -0.6696  -0.4438   2.2685
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.8556247  0.1247983  -6.856 7.08e-12 ***
## Age         -0.0197011  0.0030219  -6.519 7.06e-11 ***
## G            0.0089455  0.0006813  13.131  < 2e-16 ***
## FGPct       -0.5058787  0.4494456  -1.126 0.260351
## X3PPct       0.0740339  0.1081417   0.685 0.493596
## eFGPct       0.5427026  0.4386435   1.237 0.216002
## FTPct        0.2953804  0.0870254   3.394 0.000688 ***
## TRB         -0.0101910  0.0081698  -1.247 0.212251
## AST          0.0112233  0.0110619   1.015 0.310298
## STL         -0.0214882  0.0454635  -0.473 0.636465
## PTS         -0.0074471  0.0039993  -1.862 0.062589 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13248  on 12138  degrees of freedom
## Residual deviance: 12936  on 12128  degrees of freedom
## AIC: 12958
##
## Number of Fisher Scoring iterations: 4
```

The Probit model has only 4 significant coefficients (including the intercept), but all are highly significant. As we saw with the scatterplots, age decreases the probability of being traded. The coefficient of `FGPct` is negative, which we would expect. However, `X3Pct` and `eFGPct` both have positive effects. It appears that players who are better at scoring 3-point baskets are *more* likely to be traded away. However, the coefficient of `PTS` is negative (though not significant at the 10% level).

3.3 Logit Model

```
logit.mod <- glm(Traded ~ ., family=binomial(link="logit"), data=nba_df)
summary(logit.mod)
```

```
##
## Call:
## glm(formula = Traded ~ ., family = binomial(link = "logit"),
##     data = nba_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0361  -0.7875  -0.6686  -0.4585   2.2270
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.452552   0.220228  -6.596 4.23e-11 ***
## Age         -0.031771   0.005210  -6.098 1.07e-09 ***
## G            0.015381   0.001186  12.968  < 2e-16 ***
## FGPct       -0.860706   0.769873  -1.118  0.26357
## X3PPct       0.139790   0.186405   0.750  0.45330
## eFGPct       0.895702   0.749718   1.195  0.23220
## FTPct        0.508483   0.157358   3.231  0.00123 **
## TRB         -0.018847   0.014063  -1.340  0.18017
## AST          0.016573   0.018735   0.885  0.37638
## STL         -0.035358   0.077752  -0.455  0.64929
## PTS         -0.012750   0.006823  -1.869  0.06168 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13248  on 12138  degrees of freedom
## Residual deviance: 12941  on 12128  degrees of freedom
## AIC: 12963
```

```
##
## Number of Fisher Scoring iterations: 4
```

Unsurprisingly, the coefficient estimates of the logit model are extremely similar to the probit model in relation to each other. The signs of the coefficients are all the same, and the same coefficients are statistically significant at the 5% level. There is not much else we can examine that we didn't already remark on for the probit model.

Performance

There are too many independent variables to feasibly graph our models. To determine the best model, we turn to the standard binary classification metrics: binary cross entropy and ROC AUC (area under receiver operating characteristic curve).

```
pred <- fitted(linear.mod)
pred[pred <= 0] <- 1e-7
linear.ce <- mean(-(nba_df$Traded * log(pred) + (1 - nba_df$Traded) * log(1 - pred)))
probit.ce <- mean(-(nba_df$Traded * log(fitted(probit.mod)) +
                      (1 - nba_df$Traded) * log(1 - fitted(probit.mod))))
logit.ce <- mean(-(nba_df$Traded * log(fitted(logit.mod)) +
                     (1 - nba_df$Traded) * log(1 - fitted(logit.mod))))

ce <- c(linear.ce, probit.ce, logit.ce)
auc <- c(roc(nba_df$Traded, pred)$auc,
         roc(nba_df$Traded, fitted(probit.mod))$auc,
         roc(nba_df$Traded, fitted(logit.mod))$auc)
metrics <- cbind("Binary CE"=ce, "AUC"=auc)
rownames(metrics) <- c("Linear", "Probit", "Logit")
metrics %>% kable()
```

	Binary CE	AUC
Linear	0.8328288	0.5961851
Probit	0.5328268	0.5997310
Logit	0.5330523	0.6000069

All three models perform fairly similarly. The linear probability model has a slightly higher binary cross entropy, but performs similarly on AUC. While the AUC is not great for any of the models, all three models outperform the "baseline" of 0.5, so they are at least better than a random or majority-class guess (for some decision threshold).

Since the models are so close, and we saw in our previous analysis that the relative magnitudes of the coefficients are similar, the choice of which model to use for prediction does not make a large difference. We will choose to use the Logit model because it does not suffer from the unrestricted range problem like the linear model, and because it is more readily interpretable than the Probit model.

3.4 Prediction

First, we will create synthetic data to test some hypotheses using our model. We will evaluate 4 potential players: a young rookie with relatively little time in the league; a mid-career player with a good scoring record; a mid-career player with a good defensive record; and an established player with a strong scoring record.

```r
nba_test <- data.frame(
  "Age"=c(21, 25, 26, 32),
  "G"=c(21,54,58,56),
  "FGPct"=c(0.331,0.446,0.396,0.426),
  "X3PPct"=c(0.328,0.444,0.346,0.361),
  "eFGPct"=c(0.419,0.526,0.489,0.514),
  "FTPct"=c(0.827,0.854,0.746,0.863),
  "TRB"=c(2.1,1.9,10.2,4.1),
  "AST"=c(0.7,2.3,1.4,4.2),
  "STL"=c(0.3,0.9,1.0,0.8),
  "PTS"=c(4.6,18.1,11.1,24.8)
)
predict(logit.mod, nba_test, type="response")
```

```
##         1         2         3         4
## 0.2079697 0.2510609 0.2301326 0.2014872
```

The model prediction for all four players is around the 20-25% range. Unsurprisingly, the final player (the most experienced) has the lowest likelihood of being traded. However, the value is still relatively high at 20.1%. Somewhat surprisingly, the rookie player is only barely higher at 20.8%, with the two mid-career players coming in at 25.1% and 23%.

Our observations of the three models above indicated that higher scoring players are more likely to be traded, up to a point. This may be because there is a benefit to the team that trades away a good player and receives a good deal in return. It is possible that not all trades are done to get rid of a bad player, which makes the prediction task much more difficult. It is the average player, more than either the rookie or the experienced pro, that is more likely to be traded according to our model.

Based on the performance of our models, the predictions are relatively unreliable. It is possible that gains in performance could come from additional feature engineering, especially with our hypothesis about non-constant marginal effects outlined above.