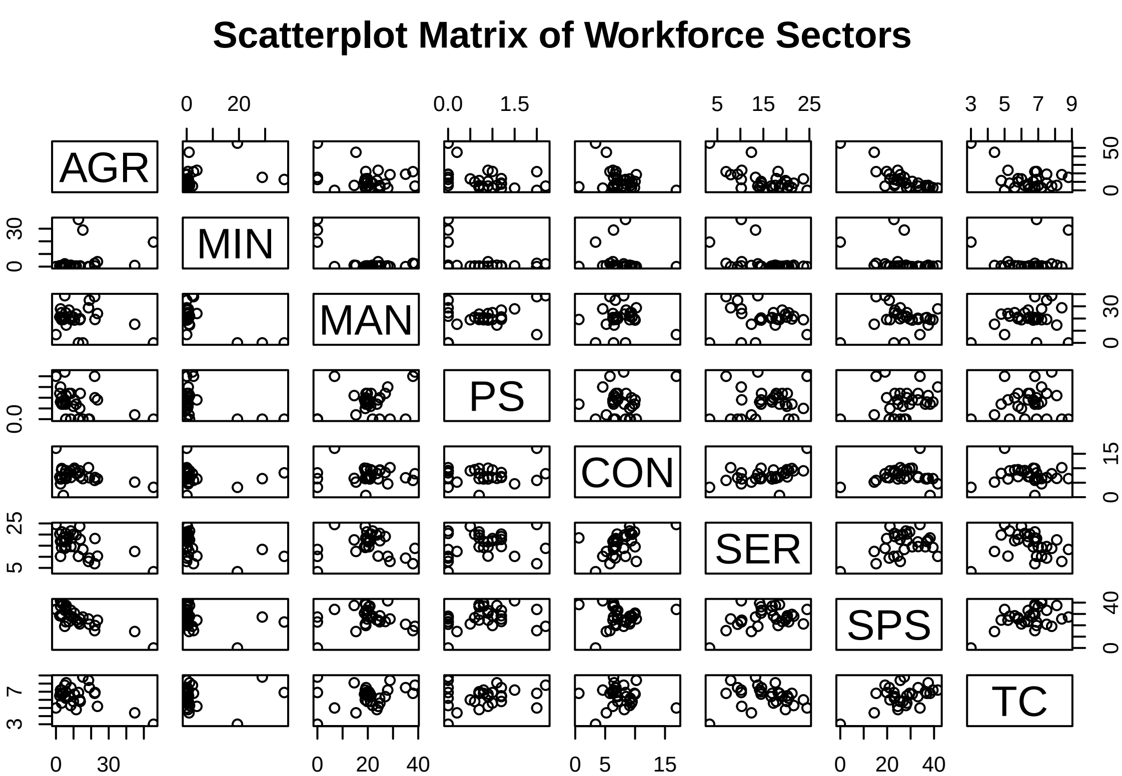
574_HW1

Matthew Stoebe

2025-04-01

e.

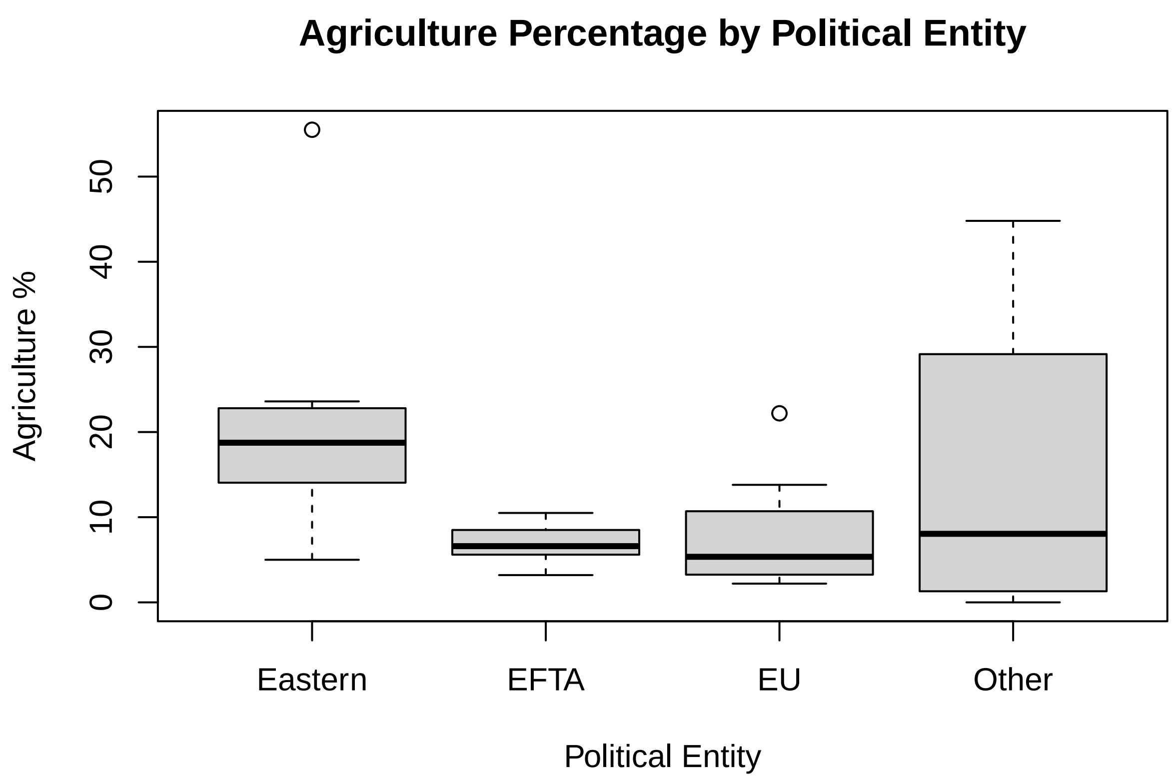
```r
workforce <- read.table("Data/workforce.txt", header=TRUE)

sectorData <- workforce[, c("AGR", "MIN", "MAN", "PS", "CON", "SER", "SPS", "TC")]

pairs(sectorData, main="Scatterplot Matrix of Workforce Sectors")
```



There are signifficant differences between each metric. Some like Min may need to be log transformed due to high outliers. Others like Con - Ser have a strong positive correlation. Some non linear patterns also emerge

b.

```r
boxplot(AGR ~ Group, data=workforce,
        main="Agriculture Percentage by Political Entity",
        xlab="Political Entity", ylab="Agriculture %")
```



We see that EU and EFTA have relatively low rate of agriculture. Eastern countries are high on averaeg and then the "other" group has a wide variance but a low median aggriculture rate.

c.

```r
summaryStats <- data.frame(
  Mean = apply(sectorData, 2, mean),
  Variance = apply(sectorData, 2, var),
  StdDev = apply(sectorData, 2, sd)
)

# Correlation matrix:
corMatrix <- cor(sectorData)
print(corMatrix)
```

```
##            AGR         MIN         MAN         PS         CON         SER
## AGR  1.0000000  0.31606875 -0.25438889 -0.3823566 -0.34861031 -0.60471243
## MIN  0.3160688  1.00000000 -0.67193466 -0.3873780 -0.12902071 -0.40654843
## MAN -0.2543889 -0.67193466  1.00000000  0.3878906 -0.03445846 -0.03294004
## PS  -0.3823566 -0.38737805  0.38789059  1.0000000  0.16479638  0.15498141
## CON -0.3486103 -0.12902071 -0.03445846  0.1647964  1.00000000  0.47308319
```

```
## SER -0.6047124 -0.40654843 -0.03294004  0.1549814  0.47308319  1.00000000
## SPS -0.8114755 -0.31641839  0.05028408  0.2377402  0.07200705  0.38798122
## TC  -0.4873331  0.04470213  0.24290323  0.1053667 -0.05460530 -0.08489430
##             SPS          TC
## AGR -0.81147553 -0.48733306
## MIN -0.31641839  0.04470213
## MAN  0.05028408  0.24290323
## PS   0.23774016  0.10536672
## CON  0.07200705 -0.05460530
## SER  0.38798122 -0.08489430
## SPS  1.00000000  0.47492344
## TC   0.47492344  1.00000000
```

```r
print(summaryStats)
```

```
##          Mean    Variance     StdDev
## AGR 12.186667 151.4598161 12.3069012
## MIN  3.446667  78.6011954  8.8657315
## MAN 20.286667  89.4308506  9.4567886
## PS   0.800000   0.3855172  0.6209003
## CON  7.530000   7.4697586  2.7330859
## SER 15.636667  26.6272299  5.1601579
## SPS 26.993333  76.2489195  8.7320627
## TC   6.453333   1.5211954  1.2333675
```

Most correlations are negative because an increase in one sector drives a decrease in others. Generally, Most correlations are negative and relatively weak.

d.

```r
library(dplyr)
```

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

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

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

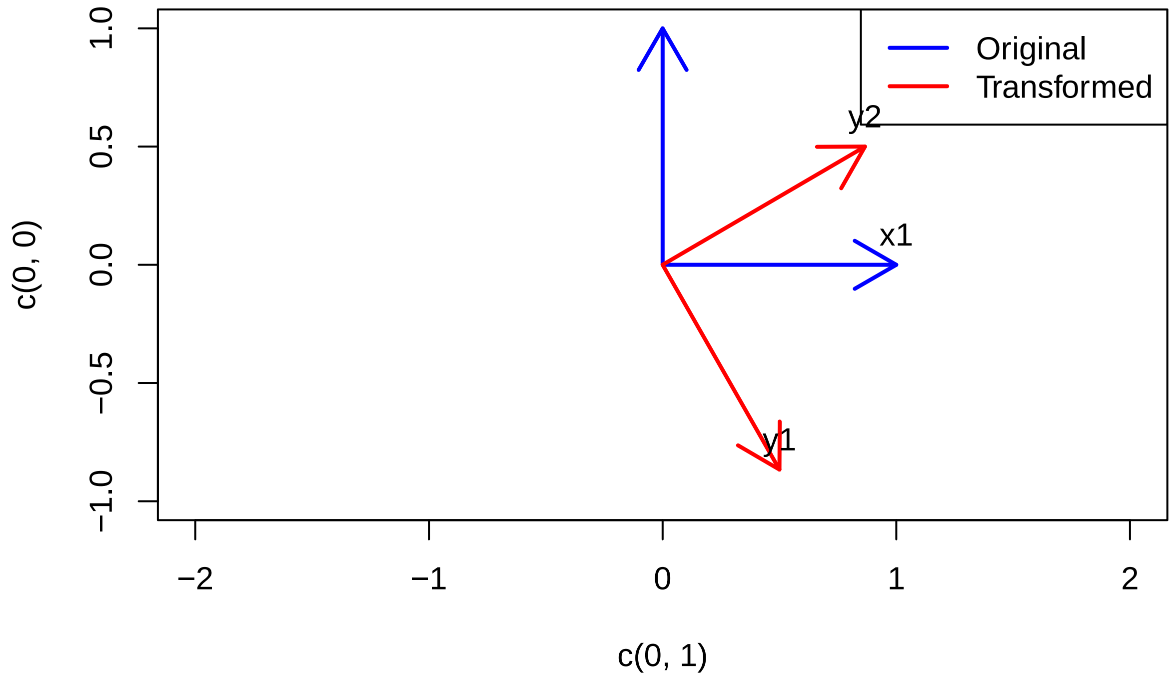
```r
groupStats <- workforce %>%
  group_by(Group) %>%
  summarise(across(c(AGR, MIN, MAN, PS, CON, SER, SPS, TC),
                   list(mean = mean, variance = var, sd = sd),
                   .names = "{.col}_{.fn}"))
print(groupStats)
```

```
## # A tibble: 4 x 25
##   Group   AGR_mean AGR_variance AGR_sd MIN_mean MIN_variance MIN_sd MAN_mean
##   <chr>      <dbl>        <dbl>  <dbl>    <dbl>        <dbl>  <dbl>    <dbl>
## 1 EFTA        6.83         6.47   2.54    0.317       0.166   0.407     20.5
## 2 EU          7.67        35.1    5.93    0.45        0.0936  0.306     21.0
## 3 Eastern    21.5        223.    14.9    11.8       217.     14.7       20.6
## 4 Other      15.2        423.    20.6     0.45        0.15    0.387     17.2
## # i 17 more variables: MAN_variance <dbl>, MAN_sd <dbl>, PS_mean <dbl>,
## #   PS_variance <dbl>, PS_sd <dbl>, CON_mean <dbl>, CON_variance <dbl>,
## #   CON_sd <dbl>, SER_mean <dbl>, SER_variance <dbl>, SER_sd <dbl>,
## #   SPS_mean <dbl>, SPS_variance <dbl>, SPS_sd <dbl>, TC_mean <dbl>,
## #   TC_variance <dbl>, TC_sd <dbl>
```

There are too many findings to summarize, but it seems that EU and EFTA have lower agriculture than the eastern, and lower mining than eastern on average. There is also higher variance in the eastern and other categories. This is all plausible.

e. It is not easy to sumarize findings especially without a specific research question. There are so many things to look at and test for that it quickly becomes overwhelming. To effectivley analyze this dataset I wouuld be curious about a specific research question and would generate plots towards that goal.

```r
x1 <- c(1, 0)
x2 <- c(0, 1)

Q <- matrix(c(1/2, sqrt(3)/2, -sqrt(3)/2, 1/2), nrow=2, byrow=TRUE)

y1 <- Q %*% x1
y2 <- Q %*% x2

plot(c(0,1), c(0,0), type="n", xlim=c(-2,2), ylim=c(-1,1))
arrows(0, 0, x1[1], x1[2], col="blue", lwd=2)
arrows(0, 0, x2[1], x2[2], col="blue", lwd=2)
text(x1[1], x1[2], "x1", pos=3)
text(x2[1], x2[2], "x2", pos=3)

arrows(0, 0, y1[1], y1[2], col="red", lwd=2)
arrows(0, 0, y2[1], y2[2], col="red", lwd=2)
text(y1[1], y1[2], "y1", pos=3)
text(y2[1], y2[2], "y2", pos=3)
legend("topright", legend=c("Original", "Transformed"), col=c("blue", "red"), lwd=2)
```


Original
Transformed
y2
x1
y1
1.0
0.5
0.0
−0.5
−1.0
c(0, 0)
−2
−1
0
1
2
c(0, 1)