Project EDA

Lydia Gibson

2022-03-27

```
stem<-read.csv(url("https://raw.githubusercontent.com/lgibson7/data/master/college-majors/women-stem.cs
options(scipen = 100) #surpress scientific notation
head(stem)
```

```
##   Rank Major_code                                    Major Major_category
## 1    1       2419                    PETROLEUM ENGINEERING    Engineering
## 2    2       2416           MINING AND MINERAL ENGINEERING    Engineering
## 3    3       2415                METALLURGICAL ENGINEERING    Engineering
## 4    4       2417 NAVAL ARCHITECTURE AND MARINE ENGINEERING    Engineering
## 5    5       2418                     NUCLEAR ENGINEERING    Engineering
## 6    6       2405                   CHEMICAL ENGINEERING    Engineering
##   Total   Men Women ShareWomen Median
## 1  2339  2057   282  0.1205643 110000
## 2   756   679    77  0.1018519  75000
## 3   856   725   131  0.1530374  73000
## 4  1258  1123   135  0.1073132  70000
## 5  2573  2200   373  0.1449670  65000
## 6 32260 21239 11021  0.3416305  65000
```

```
dim(stem)
```

```
## [1] 76  9
```

```
stem$fMajor_category<-as.factor(stem$Major_category) #set major category as a factor
head(stem) #view first 6 rows of data
```

```
##   Rank Major_code                                    Major Major_category
## 1    1       2419                    PETROLEUM ENGINEERING    Engineering
## 2    2       2416           MINING AND MINERAL ENGINEERING    Engineering
## 3    3       2415                METALLURGICAL ENGINEERING    Engineering
## 4    4       2417 NAVAL ARCHITECTURE AND MARINE ENGINEERING    Engineering
## 5    5       2418                     NUCLEAR ENGINEERING    Engineering
## 6    6       2405                   CHEMICAL ENGINEERING    Engineering
##   Total   Men Women ShareWomen Median fMajor_category
## 1  2339  2057   282  0.1205643 110000     Engineering
## 2   756   679    77  0.1018519  75000     Engineering
## 3   856   725   131  0.1530374  73000     Engineering
## 4  1258  1123   135  0.1073132  70000     Engineering
## 5  2573  2200   373  0.1449670  65000     Engineering
## 6 32260 21239 11021  0.3416305  65000     Engineering
```

```
levels(stem$fMajor_category)
```

```
## [1] "Biology & Life Science"  "Computers & Mathematics"
## [3] "Engineering"             "Health"
```

```
## [5] "Physical Sciences"
```

```
stem2<-stem[,-c(2:3)] #remove major code and major and create new dataset stem2
head(stem2) #view first 6 rows of new dataset
```

```
##   Rank Major_category Total   Men Women ShareWomen Median fMajor_category
## 1    1    Engineering  2339  2057   282  0.1205643 110000     Engineering
## 2    2    Engineering   756   679    77  0.1018519  75000     Engineering
## 3    3    Engineering   856   725   131  0.1530374  73000     Engineering
## 4    4    Engineering  1258  1123   135  0.1073132  70000     Engineering
## 5    5    Engineering  2573  2200   373  0.1449670  65000     Engineering
## 6    6    Engineering 32260 21239 11021  0.3416305  65000     Engineering
```

```
dim(stem2)
```

```
## [1] 76  8
```

```
for (i in colnames(stem2[,c(1:8)])){stem2[[i]] <- as.numeric(stem2[[i]])}
```

```
## Warning: NAs introduced by coercion
```

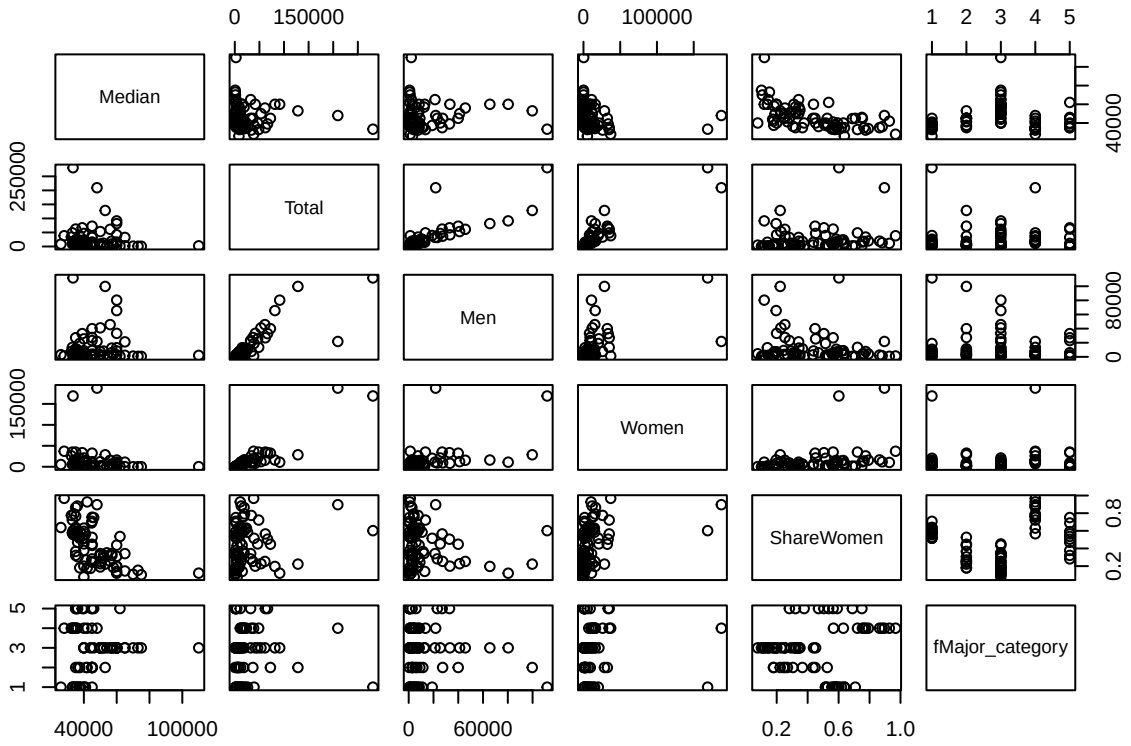
```
head(stem2)
```

```
##   Rank Major_category Total   Men Women ShareWomen Median fMajor_category
## 1    1             NA  2339  2057   282  0.1205643 110000               3
## 2    2             NA   756   679    77  0.1018519  75000               3
## 3    3             NA   856   725   131  0.1530374  73000               3
## 4    4             NA  1258  1123   135  0.1073132  70000               3
## 5    5             NA  2573  2200   373  0.1449670  65000               3
## 6    6             NA 32260 21239 11021  0.3416305  65000               3
```

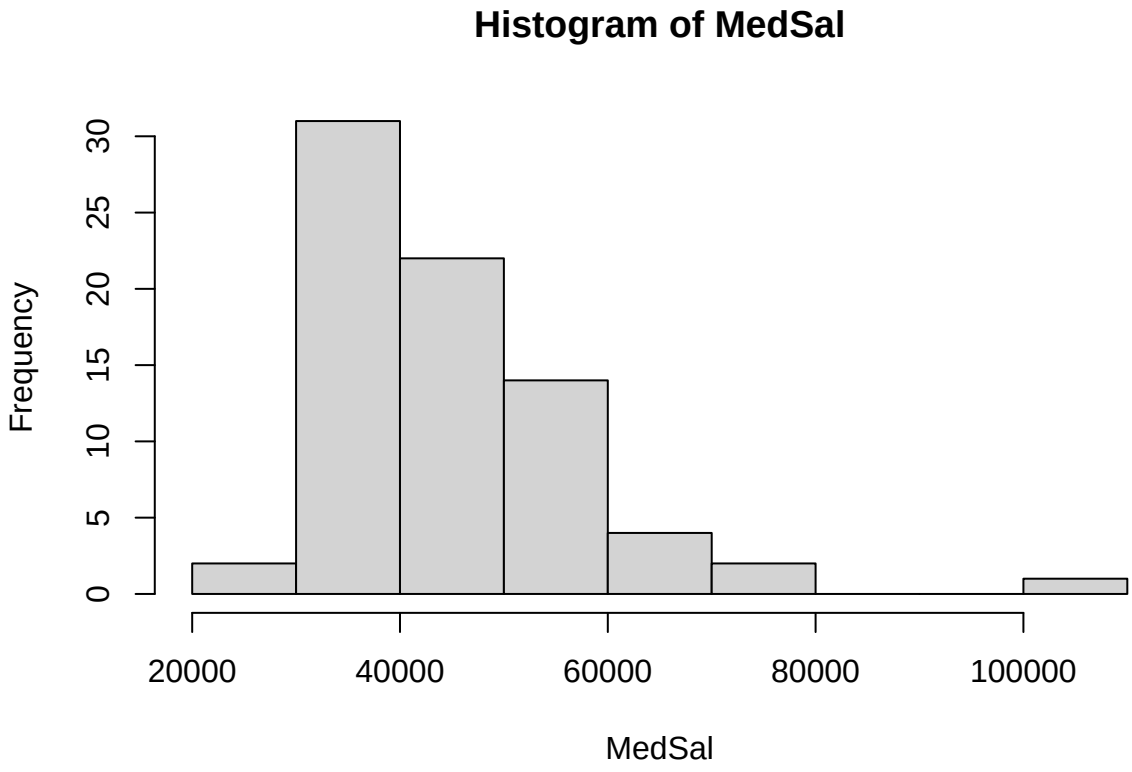
```
summary(stem2)
```

```
##       Rank       Major_category     Total             Men
##  Min.   : 1.00   Min.   : NA    Min.   :   609   Min.   :   488
##  1st Qu.:19.75   1st Qu.: NA    1st Qu.:  3782   1st Qu.:  2048
##  Median :38.50   Median : NA    Median : 11048   Median :  4583
##  Mean   :38.50   Mean   :NaN    Mean   : 25515   Mean   : 12801
##  3rd Qu.:57.25   3rd Qu.: NA    3rd Qu.: 27509   3rd Qu.: 11686
##  Max.   :76.00   Max.   : NA    Max.   :280709   Max.   :111762
##                  NA's   :76
##      Women          ShareWomen          Median       fMajor_category
##  Min.   :    77   Min.   :0.07745   Min.   : 26000   Min.   :1.000
##  1st Qu.:  1228   1st Qu.:0.24792   1st Qu.: 36150   1st Qu.:2.000
##  Median :  5218   Median :0.40587   Median : 44350   Median :3.000
##  Mean   : 12715   Mean   :0.43693   Mean   : 46118   Mean   :2.908
##  3rd Qu.: 12464   3rd Qu.:0.59180   3rd Qu.: 52250   3rd Qu.:4.000
##  Max.   :187621   Max.   :0.96800   Max.   :110000   Max.   :5.000
##
```

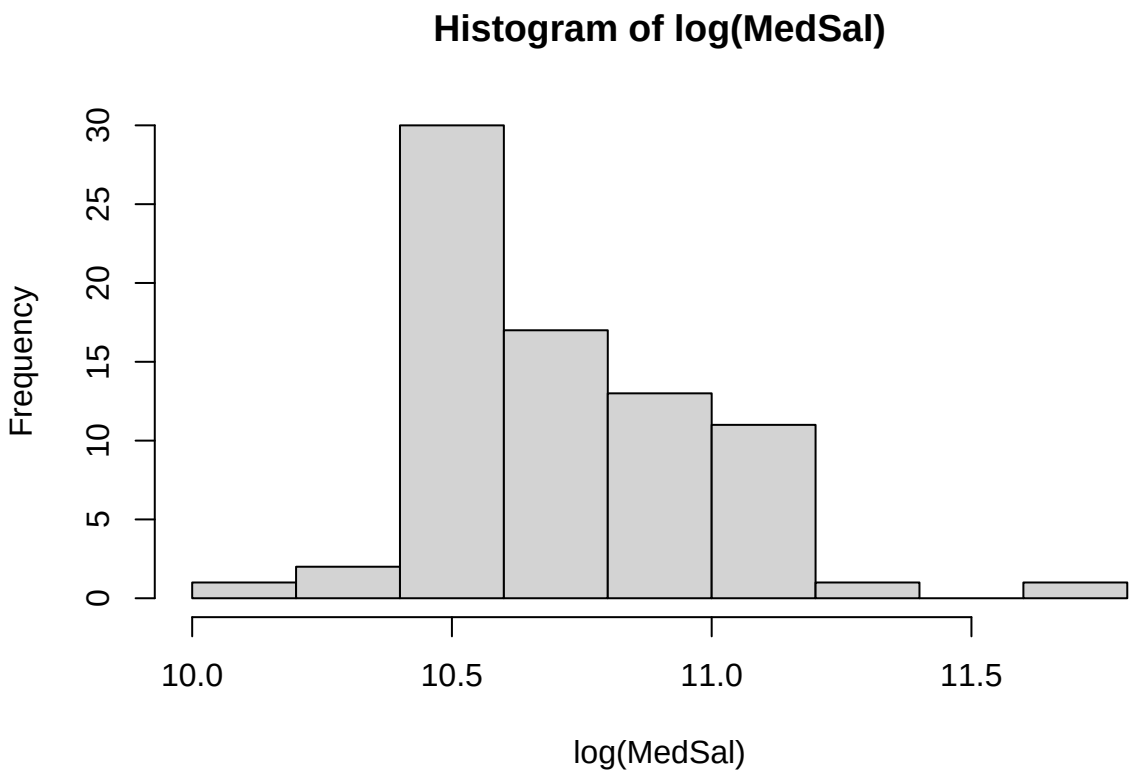
```
pairs(Median~.,data=stem2[,-c(1:2)])
```



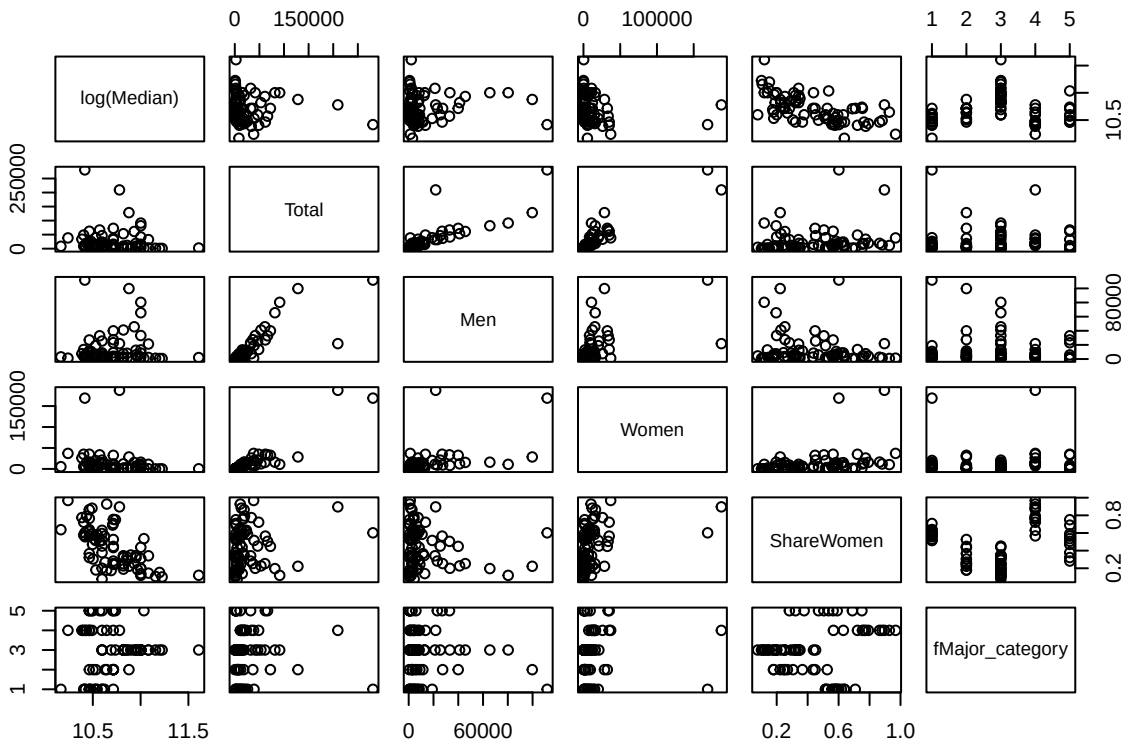

```
MedSal<-stem2$Median
hist(MedSal)
```




```
hist(log(MedSal))
```




```
pairs(log(Median)~.,data=stem2[,-c(1:2)])
```




```
lmstem<-lm(log(Median)~., data=stem2[,-c(1:2)])
summary(lmstem)
```

```
##
```

```
## Call:
## lm(formula = log(Median) ~ ., data = stem2[, -c(1:2)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40468 -0.13665  0.00585  0.11785  0.66477
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate    Std. Error t value             Pr(>|t|)
## (Intercept)     10.9227303277  0.0772973459 141.308 < 0.0000000000000002 ***
## Total            0.0000008287  0.0000010797   0.767               0.4453
## Men             -0.0000014965  0.0000022361  -0.669               0.5055
## Women                      NA            NA      NA                   NA
## ShareWomen      -0.7531337081  0.1178308531  -6.392          0.000000015 ***
## fMajor_category  0.0375574356  0.0185563575   2.024               0.0467 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1992 on 71 degrees of freedom
## Multiple R-squared:  0.422,  Adjusted R-squared:  0.3894
## F-statistic: 12.96 on 4 and 71 DF,  p-value: 0.00000005665
```

```
anova(lmstem)
```

```
## Analysis of Variance Table
##
## Response: log(Median)
##                 Df  Sum Sq Mean Sq F value        Pr(>F)
## Total            1 0.02764 0.02764  0.6966       0.40672
## Men              1 0.32314 0.32314  8.1430       0.00566 **
## ShareWomen       1 1.54332 1.54332 38.8914 0.00000002854 ***
## fMajor_category  1 0.16256 0.16256  4.0964       0.04674 *
## Residuals       71 2.81748 0.03968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmstem2<-step(lmstem)
```

```
## Start:  AIC=-240.41
## log(Median) ~ Total + Men + Women + ShareWomen + fMajor_category
##
##
## Step:  AIC=-240.41
## log(Median) ~ Total + Men + ShareWomen + fMajor_category
##
##                   Df Sum of Sq    RSS     AIC
## - Men              1   0.01777 2.8353 -241.93
## - Total            1   0.02337 2.8409 -241.78
## <none>                         2.8175 -240.41
## - fMajor_category  1   0.16256 2.9800 -238.15
## - ShareWomen       1   1.62117 4.4386 -207.87
##
## Step:  AIC=-241.93
## log(Median) ~ Total + ShareWomen + fMajor_category
##
```

```
##                   Df Sum of Sq    RSS     AIC
## - Total            1   0.00560 2.8409 -243.78
## <none>                         2.8353 -241.93
## - fMajor_category  1   0.17023 3.0055 -239.50
## - ShareWomen       1   1.96462 4.7999 -203.92
##
## Step:  AIC=-243.78
## log(Median) ~ ShareWomen + fMajor_category
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         2.8409 -243.78
## - fMajor_category  1   0.16679 3.0076 -241.45
## - ShareWomen       1   1.98335 4.8242 -205.54
```