STAT 4310 - Project

Group E: Chineze Embodi, Vaughn Jorgensen, Braxton Wilson

2023-04-26

Introduction

The Alcohol data set comes from the Woolridge package. It contains 33 variables with 9822 observations. Some of these variables include alcohol abuse, employment status, age, years of schooling, marital status, family size, and more.

Alcohol Data

The alcohol data contains several (20) categorical variables (status, married, white, exhealth, vghealth, goodhealth, fairhealth, northeast, midwest, south, centcity, outercity, qrt1, qrt2, qrt3, mothalc, fathalc, livealc, inwf, employ), but only status will be converted to a factor since all other categorical variables only have two levels. There are no missing values in any of the columns.

Variable	Description
abuse	= 1 if abuse alcohol
status	out of workforce = 1; unemployed = 2, employed = 3
unemrate	state unemployment rate
age	age in years
educ	years of schooling
married	= 1 if married
famsize	family size
white	= 1 if white
exhealth	= 1 if in excellent health
vghealth	= 1 if in very good health
goodhealth	= 1 if in good health
fairhealth	= 1 if in fair health
northeast	= 1 if live in northeast
midwest	= 1 if live in midwest
south	= 1 if live in south
centcity	= 1 if live in central city of MSA
outercity	= 1 if in outer city of MSA
qrt1	= 1 if interviewed in first quarter
qrt2	= 1 if interviewed in second quarter
qrt3	= 1 if interviewed in third quarter
beertax	state excise tax, $ per gallon
cigtax	state cigarette tax, cents per pack
ethanol	state per-capita ethanol consumption
mothalc	= 1 if mother an alcoholic
fathalc	= 1 if father an alcoholic
livealc	= 1 if lived with alcoholic
inwf	= 1 if status > 1
employ	= 1 if employed
agesq	age^2
beertaxsq	beertax^2
cigtaxsq	cigtax^2
ethanolsq	ethanol^2
educsq	educ^2

```
##      abuse     status   unemrate        age       educ    married    famsize
##          0          0          0          0          0          0          0
##      white   exhealth   vghealth goodhealth fairhealth  northeast    midwest
##          0          0          0          0          0          0          0
##      south   centcity  outercity       qrt1       qrt2       qrt3    beertax
##          0          0          0          0          0          0          0
##     cigtax    ethanol    mothalc    fathalc    livealc       inwf     employ
##          0          0          0          0          0          0          0
##      agesq  beertaxsq   cigtaxsq  ethanolsq     educsq
##          0          0          0          0          0
```

```
## 'data.frame':    9822 obs. of  33 variables:
##  $ abuse     : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ status    : int  1 3 3 3 3 3 3 1 1 3 ...
##  $ unemrate  : num  4 4 4 3.3 3.3 ...
##  $ age       : int  50 37 53 59 43 38 34 45 47 31 ...
##  $ educ      : int  4 12 9 11 10 10 10 2 5 12 ...
##  $ married   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ famsize   : int  1 5 3 1 1 1 4 2 2 1 ...
##  $ white     : int  1 1 1 1 1 1 1 1 0 1 ...
##  $ exhealth  : int  0 0 1 1 1 1 0 0 0 1 ...
##  $ vghealth  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ goodhealth: int  0 1 0 0 0 0 1 0 0 0 ...
##  $ fairhealth: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ northeast : int  0 0 0 1 1 1 0 0 0 0 ...
##  $ midwest   : int  1 1 1 0 0 0 1 1 1 1 ...
##  $ south     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ centcity  : int  0 0 0 1 1 1 0 0 0 1 ...
##  $ outercity : int  0 0 0 0 0 0 1 1 1 0 ...
##  $ qrt1      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ qrt2      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ qrt3      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ beertax   : num  0.334 0.334 0.334 0.24 0.24 ...
##  $ cigtax    : num  38 38 38 26 26 26 20 20 20 20 ...
##  $ ethanol   : num  2.04 2.04 2.04 2.45 2.45 ...
##  $ mothalc   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ fathalc   : int  0 0 0 0 1 0 1 0 1 0 ...
##  $ livealc   : int  0 0 0 0 1 0 1 0 1 0 ...
##  $ inwf      : int  0 1 1 1 1 1 1 0 0 1 ...
##  $ employ    : int  0 1 1 1 1 1 1 0 0 1 ...
##  $ agesq     : int  2500 1369 2809 3481 1849 1444 1156 2025 2209 961 ...
##  $ beertaxsq : num  0.1116 0.1116 0.1116 0.0576 0.0576 ...
##  $ cigtaxsq  : num  1444 1444 1444 676 676 ...
##  $ ethanolsq : num  4.16 4.16 4.16 6 6 ...
##  $ educsq    : int  16 144 81 121 100 100 100 4 25 144 ...
##  - attr(*, "time.stamp")= chr "22 Jan 2013 14:09"
```

```
##   abuse status unemrate age educ married famsize white exhealth vghealth
## 1     1      1      4.0  50    4       1       1     1        0        0
## 2     0      3      4.0  37   12       1       5     1        0        0
## 3     0      3      4.0  53    9       1       3     1        1        0
## 4     0      3      3.3  59   11       1       1     1        1        0
## 5     0      3      3.3  43   10       1       1     1        1        0
## 6     0      3      3.3  38   10       1       1     1        1        0
```

```
##   goodhealth fairhealth northeast midwest south centcity outercity qrt1 qrt2
## 1          0          0         0       1     0        0         0    1    0
## 2          1          0         0       1     0        0         0    1    0
## 3          0          0         0       1     0        0         0    1    0
## 4          0          0         1       0     0        1         0    1    0
## 5          0          0         1       0     0        1         0    1    0
## 6          0          0         1       0     0        1         0    1    0
##   qrt3 beertax cigtax ethanol mothalc fathalc livealc inwf employ
## 1    0   0.334     38 2.03946       0       0       0    0      0
## 2    0   0.334     38 2.03946       0       0       0    1      1
## 3    0   0.334     38 2.03946       0       0       0    1      1
## 4    0   0.240     26 2.44998       0       0       0    1      1
## 5    0   0.240     26 2.44998       0       1       1    1      1
## 6    0   0.240     26 2.44998       0       0       0    1      1
```

```r
table(alcohol$abuse) # count for yes/1 vs no/0 alcohol abuse
```

```
##
##    0    1
## 8848  974
```

In our data set, we can see that there are 8848 observations of individuals who do not abuse alcohol (0) and 974 observations of alcohol abusers (1). This severe class imbalance may prove to be an issue.

Regression Analysis

In our regression analysis, we will be using the abuse variable as our response and all others as predictors (excluding squared variables). The abuse variable will tell is if alcohol is abused (1) or if alcohol is not abused (0). Abuse was chosen as the response because it stands out as the best option based on the other variables in this data set. Additionally, it would be interesting to see which variables are determining factors in alcohol abuse.

We will be utilizing the generalized linear model with a binomial distribution. This way, we can perform logistic regression on our categorical response variable.

```r
model <- glm(abuse ~ ., family = "binomial", data = alcohol)
summary(model)
```

```
##
## Call:
## glm(formula = abuse ~ ., family = "binomial", data = alcohol)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0141  -0.4888  -0.4254  -0.3615   2.6783
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.420181   0.454565  -5.324 1.01e-07 ***
## status2      0.191808   0.214804   0.893 0.371887
## status3     -0.075773   0.139917  -0.542 0.588124
## unemrate     0.008096   0.028565   0.283 0.776848
## age          0.000825   0.003777   0.218 0.827087
## educ        -0.038945   0.012388  -3.144 0.001667 **
## married     -0.045544   0.096912  -0.470 0.638387
## famsize     -0.153430   0.026990  -5.685 1.31e-08 ***
## white        0.272474   0.105931   2.572 0.010106 *
## exhealth    -0.271481   0.236037  -1.150 0.250077
## vghealth    -0.029678   0.235612  -0.126 0.899764
## goodhealth   0.016736   0.233943   0.072 0.942970
## fairhealth   0.047103   0.251935   0.187 0.851687
## northeast    0.093323   0.127385   0.733 0.463796
## midwest     -0.003860   0.114180  -0.034 0.973034
## south        0.012589   0.121142   0.104 0.917236
## centcity     0.204109   0.099176   2.058 0.039584 *
## outercity    0.059293   0.095178   0.623 0.533306
## qrt1         0.015503   0.095551   0.162 0.871114
## qrt2         0.031281   0.095386   0.328 0.742953
## qrt3        -0.084400   0.098848  -0.854 0.393196
## beertax      0.023020   0.104006   0.221 0.824832
## cigtax       0.006193   0.005463   1.134 0.256956
## ethanol      0.330890   0.101729   3.253 0.001143 **
## mothalc      0.408392   0.159932   2.554 0.010664 *
## fathalc      0.512149   0.140033   3.657 0.000255 ***
## livealc     -0.049249   0.141761  -0.347 0.728285
## inwf               NA         NA      NA       NA
## employ             NA         NA      NA       NA
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6349.8  on 9821  degrees of freedom
## Residual deviance: 6194.1  on 9795  degrees of freedom
## AIC: 6248.1
##
## Number of Fisher Scoring iterations: 5
```

```r
rsefull <- sqrt(deviance(model)/df.residual(model))
rsefull
```

```
## [1] 0.7952217
```

```r
# McFadden's R^2 -- Excellent fit considered to be 0.2-0.4
rsquafull <- with(summary(model), 1- deviance/null.deviance)
rsquafull
```

```
## [1] 0.02451733
```

In our full model, the variables that are statistically significant to our model at the 5% significance level are educ, famsize, white, centcity, ethanol, mothalc, and fathalc. Note the AIC score of 6248.1 which we will compare to our other models. The standard error using deviance and df comes out to 0.795. McFadden's R^2 comes out to 0.025 (good fit considered to be 0.2-0.4).

```r
modred <- step(model, trace = 0)
summary(modred)
```

```
##
## Call:
## glm(formula = abuse ~ educ + famsize + white + exhealth + centcity +
##     cigtax + ethanol + mothalc + fathalc, family = "binomial",
##     data = alcohol)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9880  -0.4900  -0.4266  -0.3626   2.7110
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.323924   0.265743  -8.745  < 2e-16 ***
## educ        -0.041300   0.011894  -3.472 0.000516 ***
## famsize     -0.161515   0.024398  -6.620 3.59e-11 ***
## white        0.246570   0.104086   2.369 0.017841 *
## exhealth    -0.271428   0.072323  -3.753 0.000175 ***
## centcity     0.166981   0.072975   2.288 0.022126 *
## cigtax       0.007433   0.004733   1.570 0.116325
## ethanol      0.327681   0.086510   3.788 0.000152 ***
## mothalc      0.384191   0.147295   2.608 0.009099 **
## fathalc      0.475994   0.084182   5.654 1.56e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6349.8  on 9821  degrees of freedom
## Residual deviance: 6200.7  on 9812  degrees of freedom
## AIC: 6220.7
##
## Number of Fisher Scoring iterations: 5
```

```
# RSE
rse <- sqrt(deviance(modred)/df.residual(modred))
rse
```

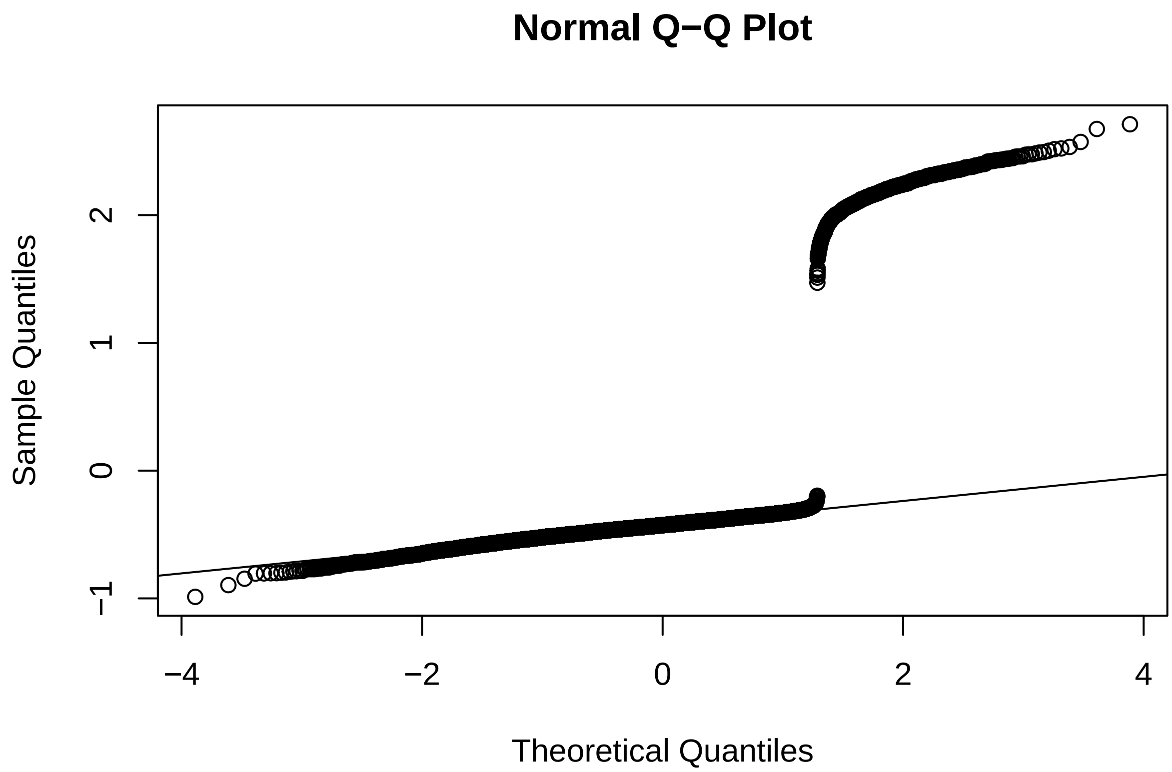
```
## [1] 0.7949535
```

```
# McFadden's R^2 -- Excellent fit considered to be 0.2-0.4
rsqua <- with(summary(modred), 1- deviance/null.deviance)
rsqua
```

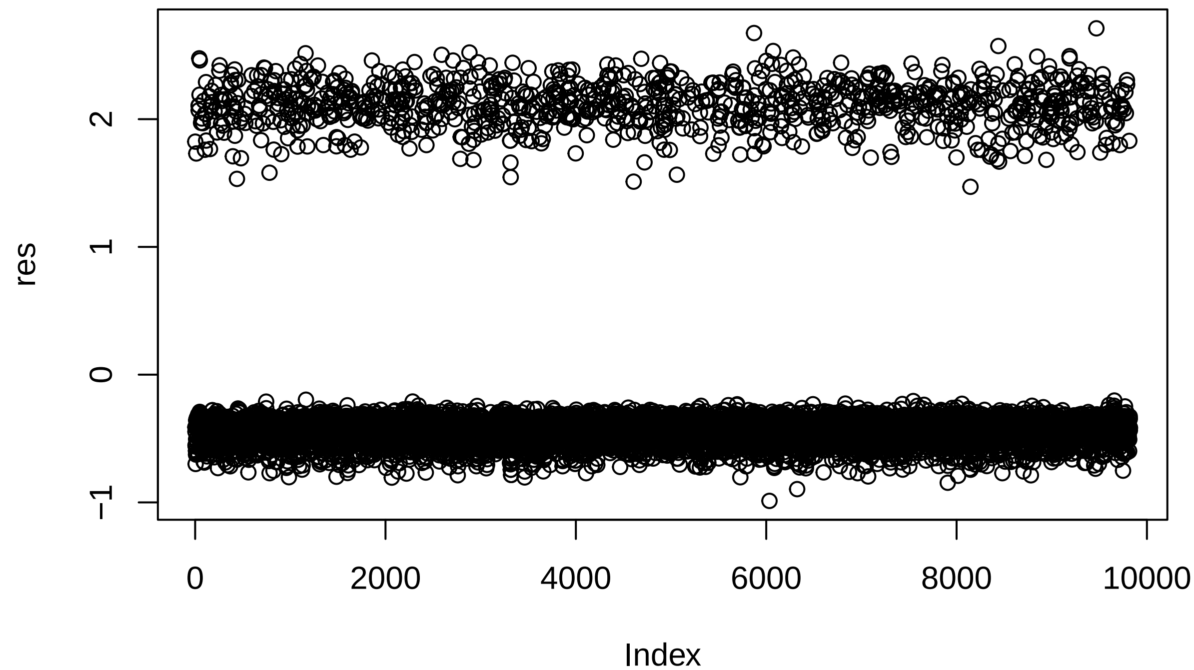
```
## [1] 0.02348328
```

After reducing our model, we went from 28 variables to 9 variables. This leaves us with the variables educ, famsize, white, exhealth, centcity, cigtax, ethanol, mothalc, and fathalc. The AIC dropped from 6248.1 to 6220.7 which may mostly be due to a decrease in variables. The RSE remained the same (0.795) but the McFadden's R^2 actually decreased to 0.023.

```
# residuals
res <- residuals(modred, type = "deviance")
qqnorm(res) # Should have normal distribution if model fits
qqline(res) # they do not...
```

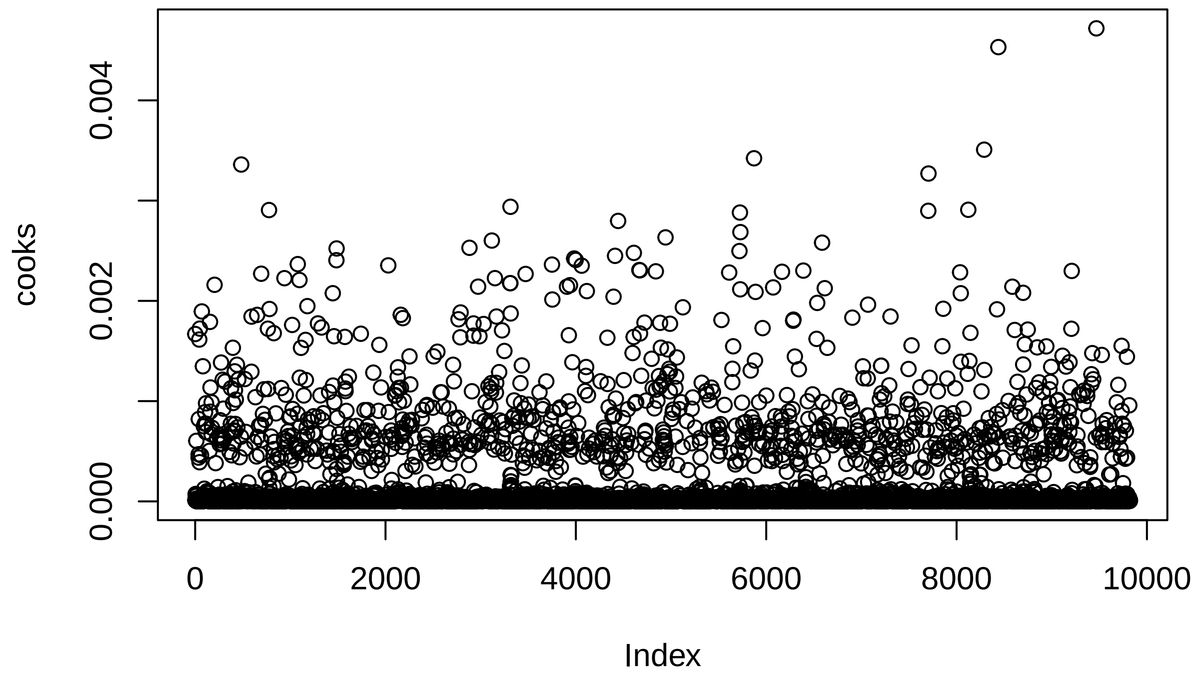



```
plot(res) # should be around 0--if not they may be outliers
```



The residuals of our model should have a normal distribution if our model fits; however, we can see that based on our qqplot that our model is not adequate. Additionally, we can see in our residual plot that many of our values are not close to 0 which promotes our models inadequacy and may suggest outliers. Let's investigate using cooks distance.

```
# cooks distance
cooks <- cooks.distance(modred)
plot(cooks) # lots of outliers
```




```
summary(cooks)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## 1.378e-06 4.467e-06 7.933e-06 1.021e-04 1.846e-05 4.718e-03
```

```
table(alcohol[cooks < 1.846e-05, "abuse"])
```

```
##
##    0
## 7366
```

```
table(alcohol[cooks < 0.00085, "abuse"])
```

```
##
##    0    1
## 8848  606
```

```
model2 <- glm(abuse ~ ., family = "binomial", data = subset(alcohol, cooks < 0.00085))
# 1.846e-05 from Q3--could not converge so chose number low enough to not get warning
modred2 <- step(model2, trace = 0)
summary(modred2)
```

```
##
```

```
## Call:
## glm(formula = abuse ~ unemrate + educ + married + famsize + white +
##     exhealth + northeast + ethanol + mothalc, family = "binomial",
##     data = subset(alcohol, cooks < 0.00085))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8680  -0.4307  -0.3337  -0.1880   3.7079
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.88316    1.09454  -7.202 5.92e-13 ***
## unemrate      0.11936    0.03476   3.434 0.000596 ***
## educ         -0.03161    0.01525  -2.073 0.038145 *
## married       0.22234    0.11531   1.928 0.053836 .
## famsize      -0.36371    0.03822  -9.517  < 2e-16 ***
## white         4.81985    1.00174   4.811 1.50e-06 ***
## exhealth     -0.46471    0.09207  -5.047 4.48e-07 ***
## northeast     0.49209    0.11814   4.165 3.11e-05 ***
## ethanol       0.54444    0.12159   4.478 7.55e-06 ***
## mothalc     -14.89815  206.44642  -0.072 0.942471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4502.0  on 9453  degrees of freedom
## Residual deviance: 4101.8  on 9444  degrees of freedom
## AIC: 4121.8
##
## Number of Fisher Scoring iterations: 16
```

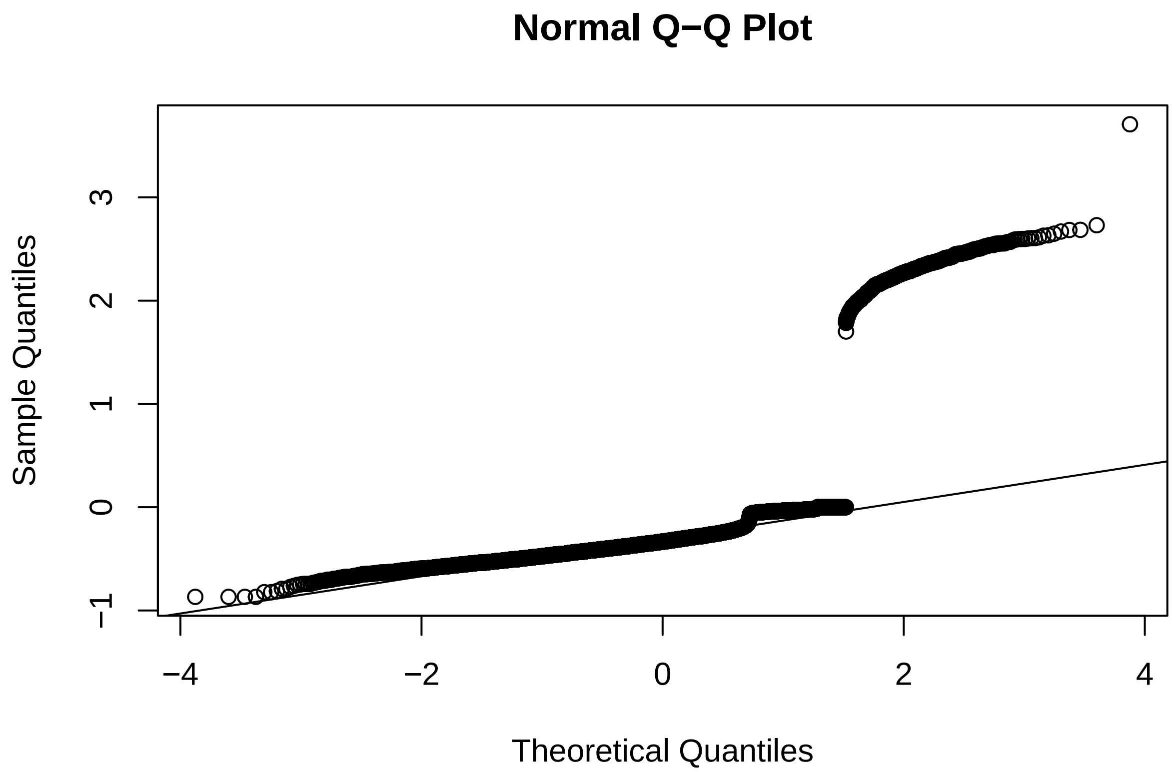
```
# RSE
rse.cook <- sqrt(deviance(modred2)/df.residual(modred2))
rse.cook
```

```
## [1] 0.6590371
```

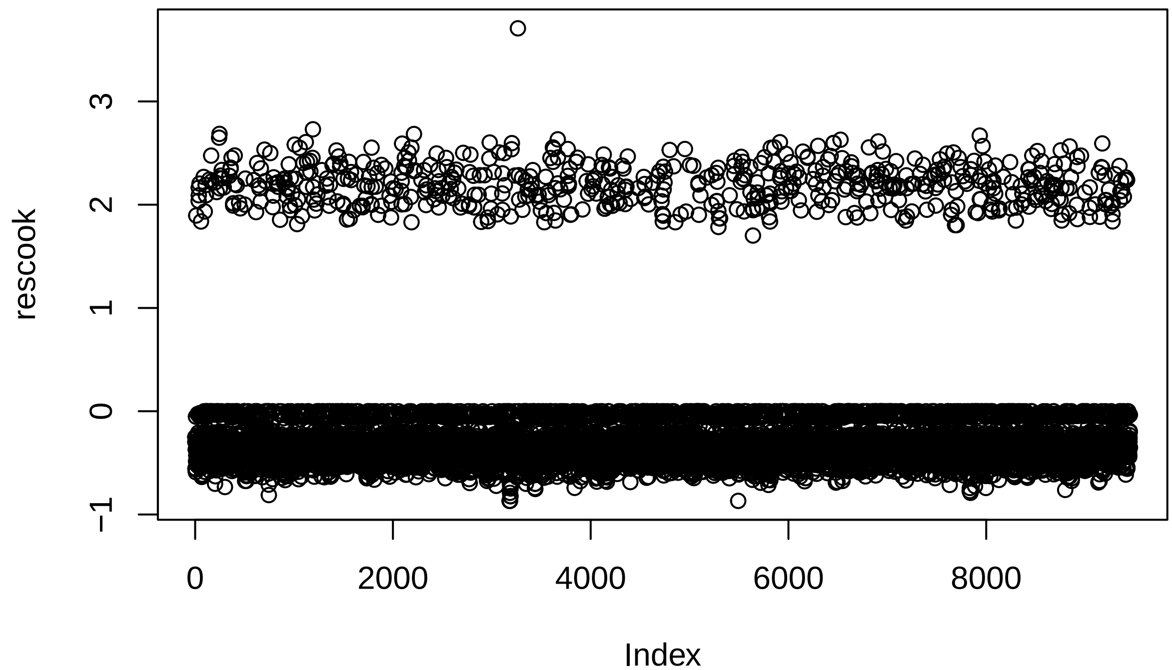
```
# R^2
rsqua.cook <- with(summary(modred2), 1- deviance/null.deviance)
rsqua.cook # 0.023 => 0.089 -- slight improvement
```

```
## [1] 0.0888996
```

```
rescook <- residuals(modred2, type = "deviance")
qqnorm(rescook) # Should have normal distribution if model fits
qqline(rescook) # they do not...
```




```
plot(rescook) # should be around 0
```

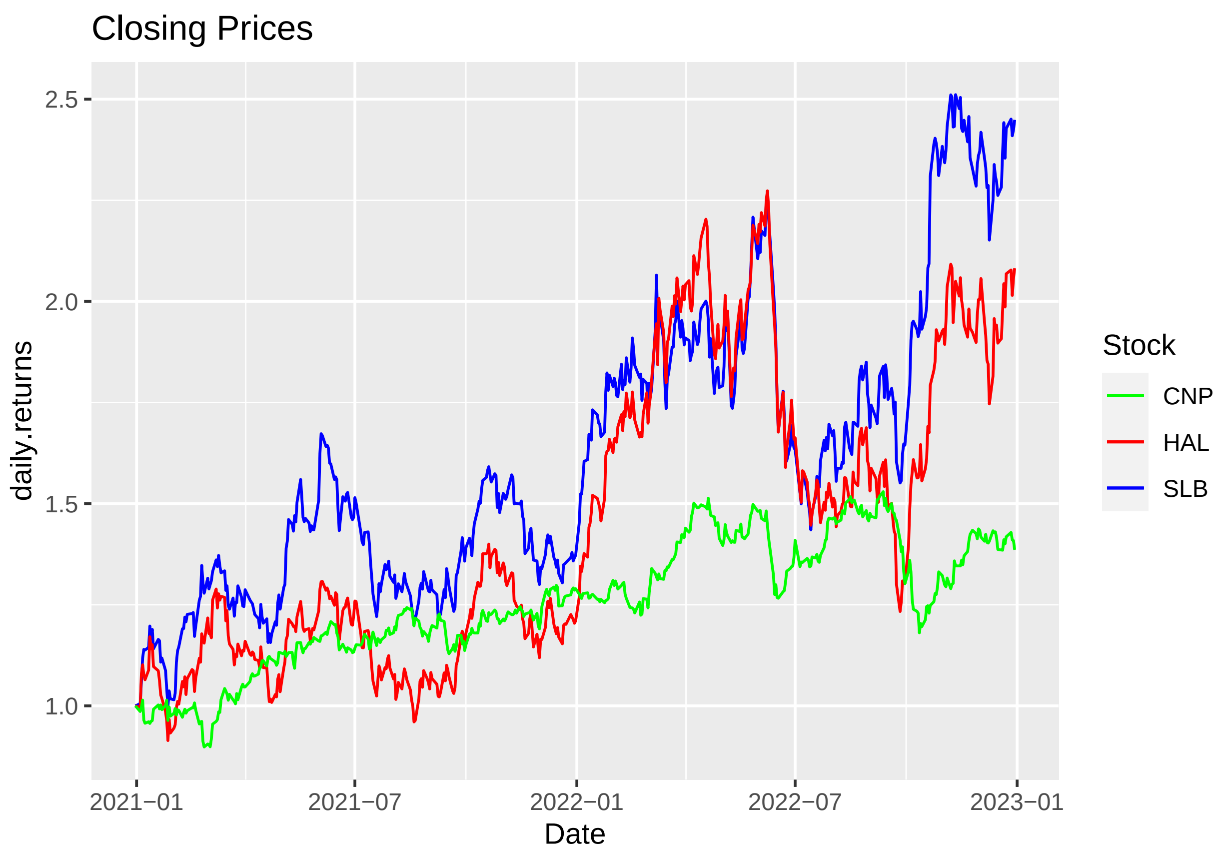


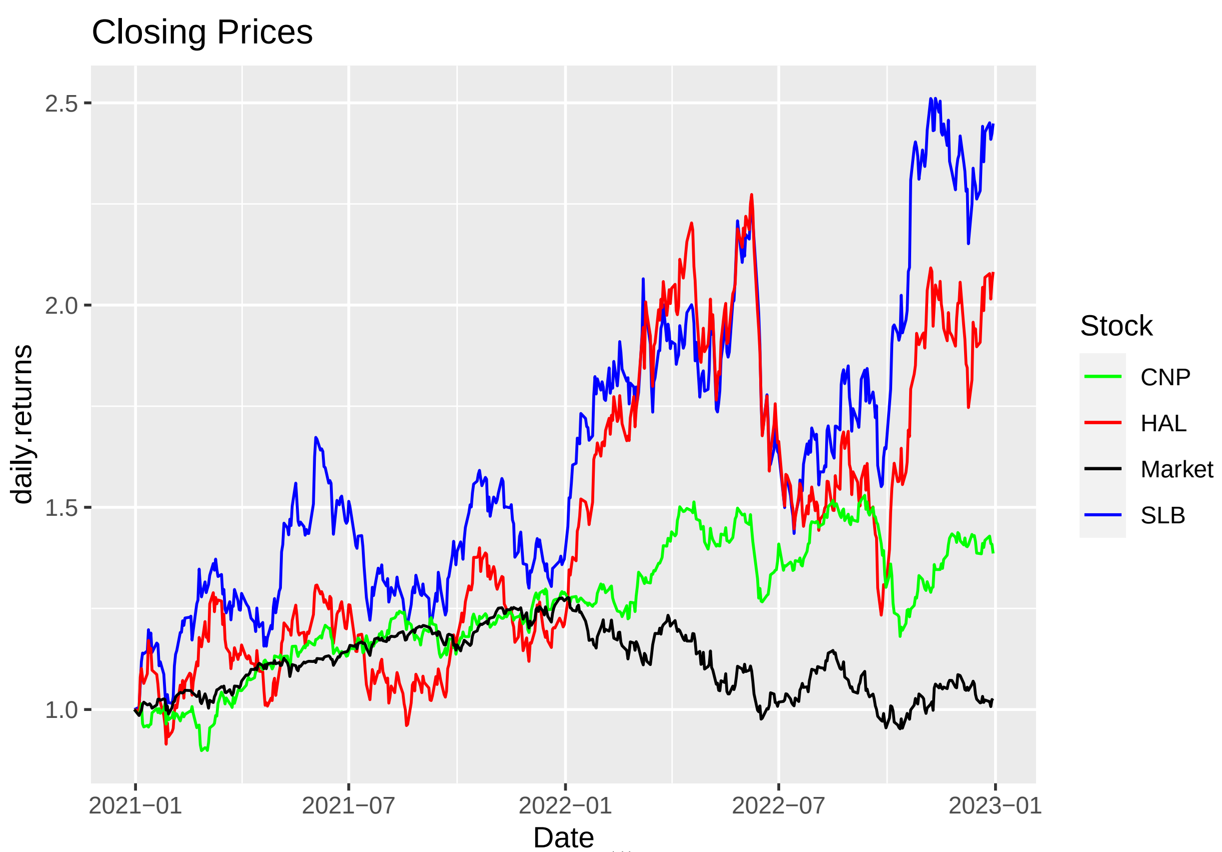
Using cooks distance, our models AIC dropped even more from 6220.7 to 4121.8. The RSE also decreased from 0.795 to 0.659, and the McFadden's R^2 increased from 0.023 to 0.089. When investigating the residuals, they still are not normal or around 0; however, it is an improvement. We could definitely improve our model further if there were more abuse = 1 observations in our data set in order to get a model to converge with a smaller subset.

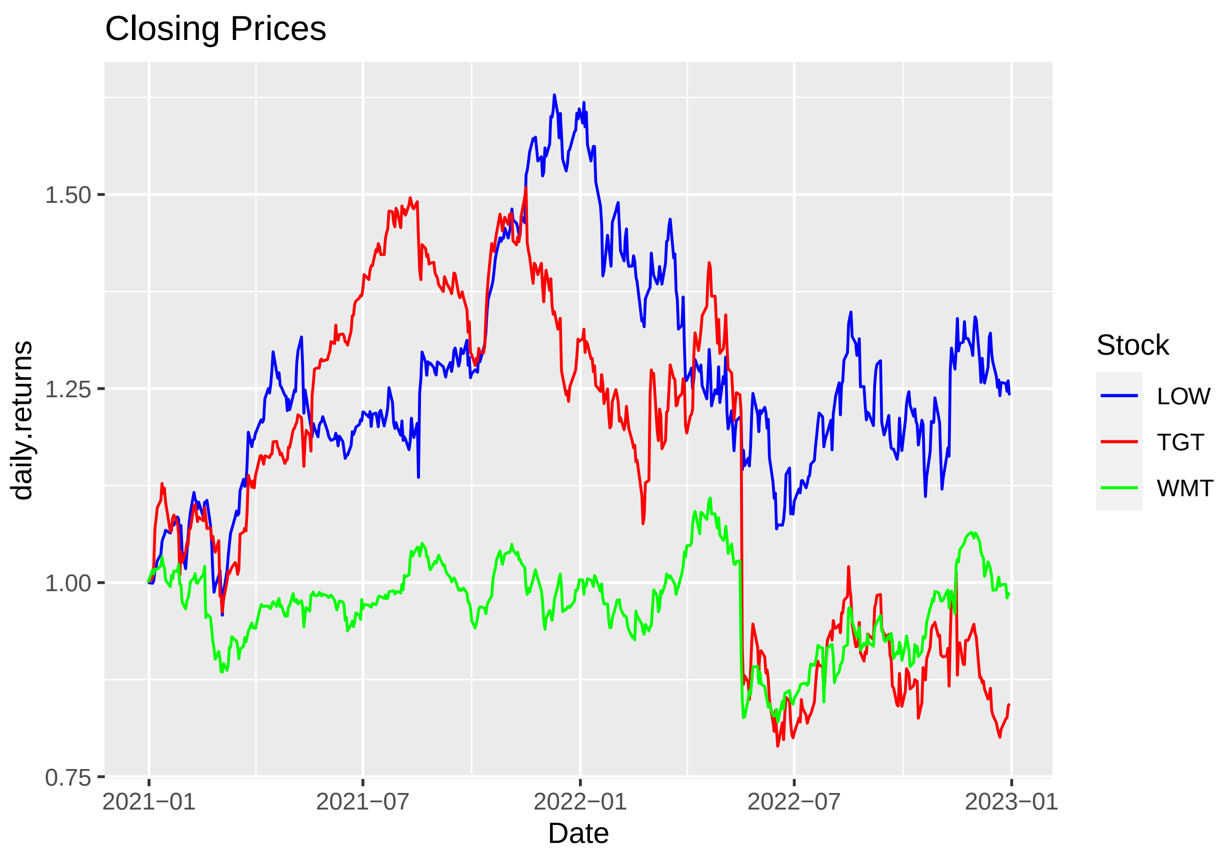
```
## fitted values in probabilities
fitted.values <- modred2$fitted.values

## prediction using 0,1
fitted <- ifelse(fitted.values >.5, 1,0)

table(alcohol[cooks < 0.00085, "abuse"], fitted)
```

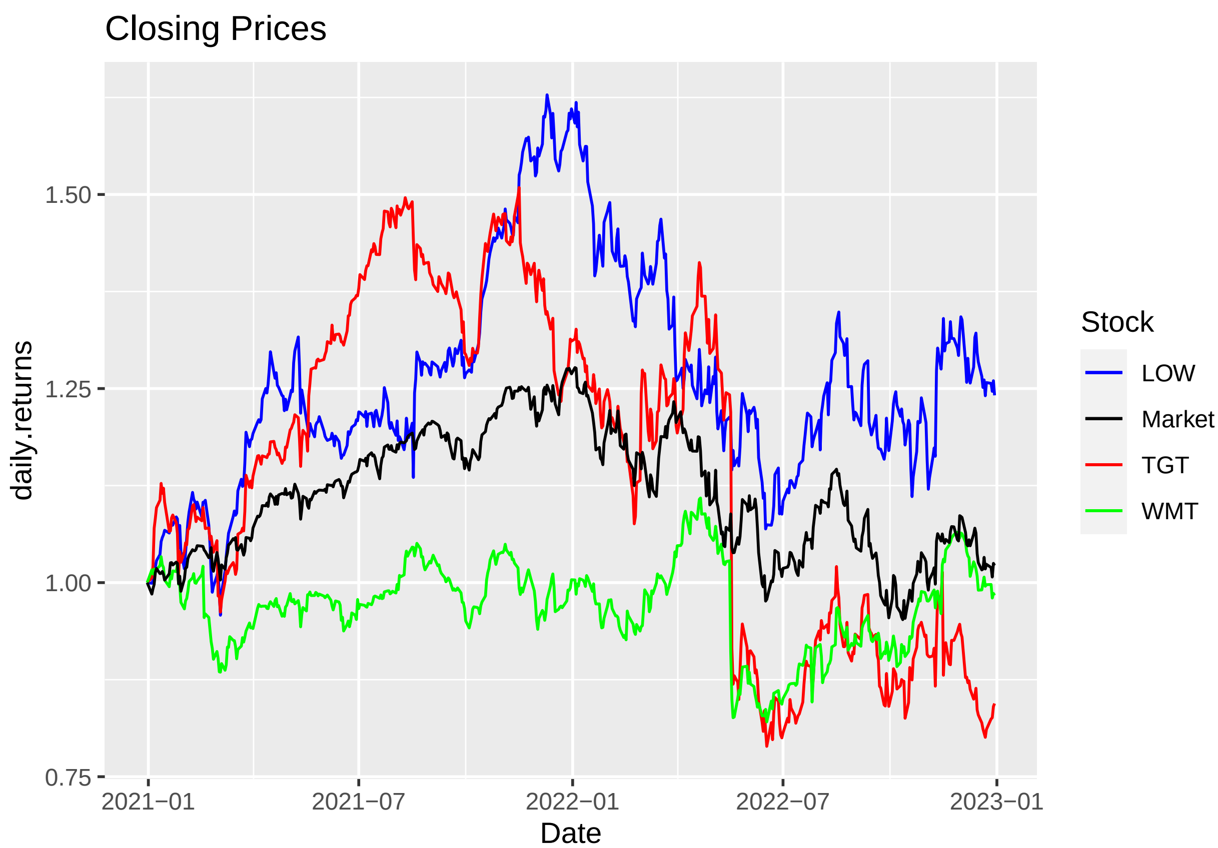
```
##    fitted
##        0
##   0 8848
##   1  606
```

Stock Analysis

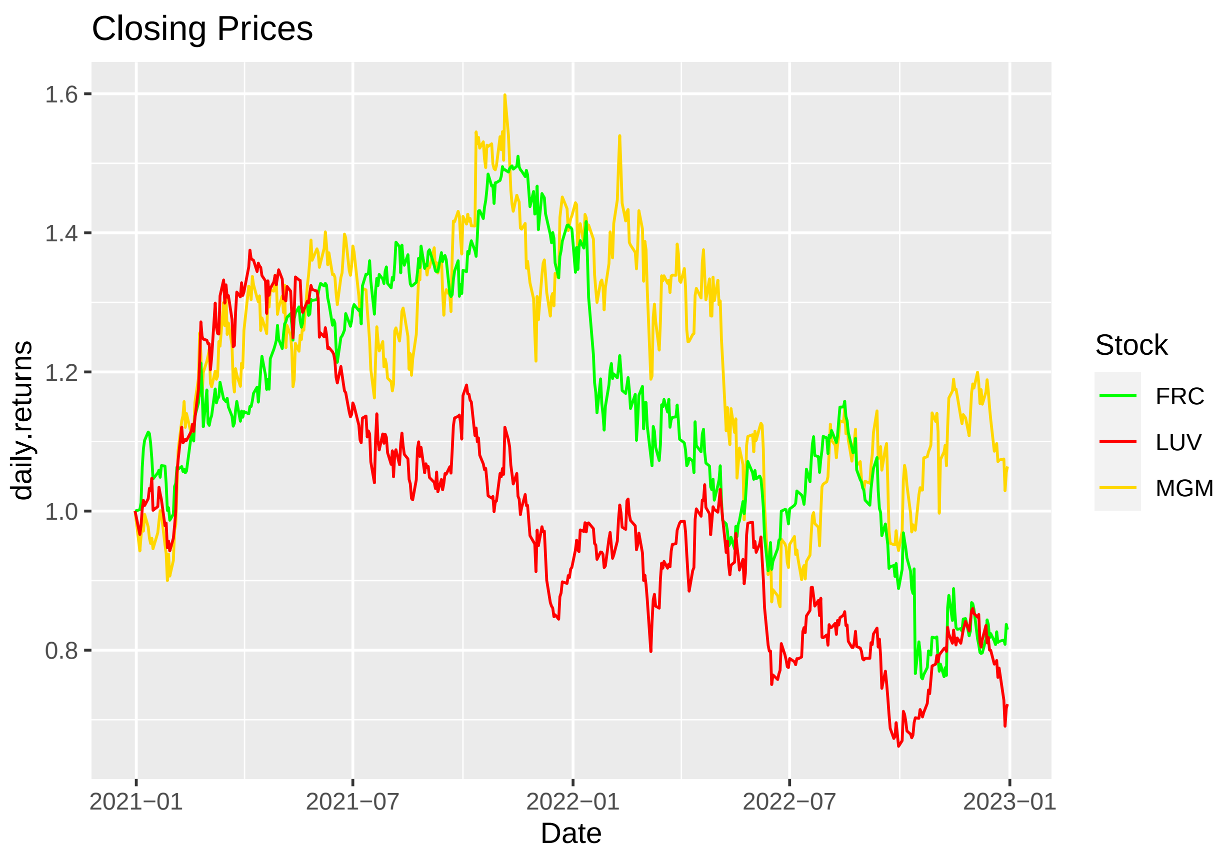




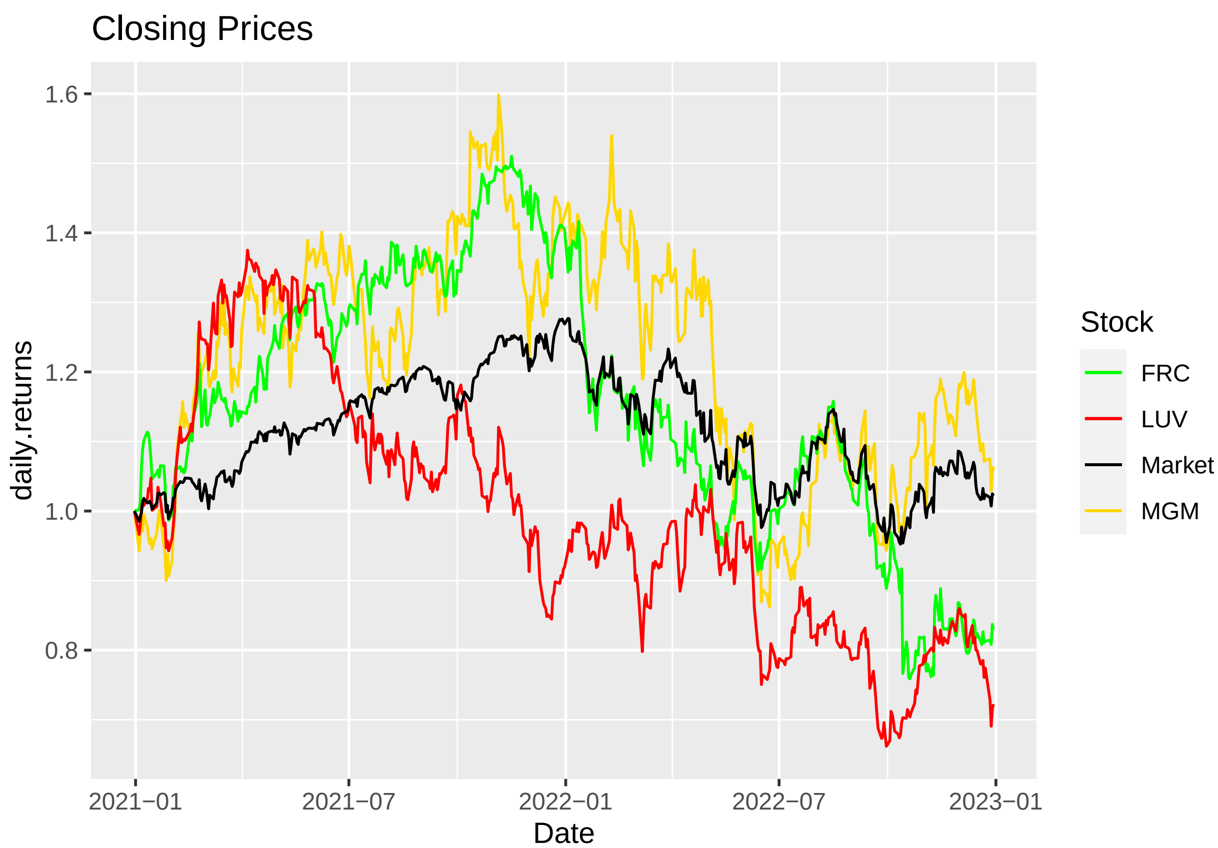

Closing Prices
daily.returns
1.50
1.25
1.00
0.75
2021−01
2021−07
2022−01
2022−07
2023−01
Date
Stock
LOW
TGT
WMT



Closing Prices
daily.returns
1.50
1.25
1.00
0.75
2021−01
2021−07
2022−01
2022−07
2023−01
Date
Stock
LOW
Market
TGT
WMT



Closing Prices
daily.returns
1.6
1.4
1.2
1.0
0.8
2021−01
2021−07
2022−01
2022−07
2023−01
Date
Stock
FRC
LUV
MGM



Closing Prices
1.6
1.4
1.2
1.0
0.8
daily.returns
2021−01
2021−07
2022−01
2022−07
2023−01
Date
Stock
FRC
LUV
Market
MGM

Company	Symbol	Return	Risk	Beta	PE
Sclumberger Ltd.	SLB	55.16%	45.54%	0.81	22.02
Halliburton Company	HAL	48.12%	47.86%	0.94	19.72
CenterPoint Energy Inc.	CNP	18.85%	22.48%	0.66	19.19
Lowe's Companies, Inc.	LOW	15.03%	29.08%	0.95	20.75
Target Corporation	TGT	-1.44%	36.58%	1.08	27.12
Walmart Inc.	WMT	1.68%	22.24%	0.51	35.53
MGM Resorts International	MGM	13.04%	44.57%	1.37	12.86
First Republic Bank	FRC	-3.28%	34.69%	1.24	1.68
Southwest Airlines Co.	LUV	-9.99%	35.45%	1.02	36.95

Out of SLB, HAL, and CNP, SLB has the highest annualized expected return and a lower risk than HAL. CNP has the lowest return and the lowest risk. HAL is the most volatile out of the three stocks. SLB also has the highest PE Ratio compared to HAL and CNP.

Between LOW, TGT, and WMT, LOW has the highest annualized expected return and the second lowest risk. TGT has a negative return but is the most volatile and has the second highest PE Ratio. WMT has the second lowest return, but has the lowest risk and highest PE Ratio out of the three stocks.

When comparing last three stocks, MGM has the highest annualized expected return compared to FRC and LUV who both have negative returns; however, MGM does have the highest risk out of the three but is also the most volatile. LUV has the lowest return and beta but has the highest PE ratio.

When comparing all nine stocks, SLB has the highest annualized expected return and the second highest annualized expected risk. LUV has the lowest return but the highest PE ratio. MGM is the most volatile and WMT is the least volatile.