problem_set5

2.

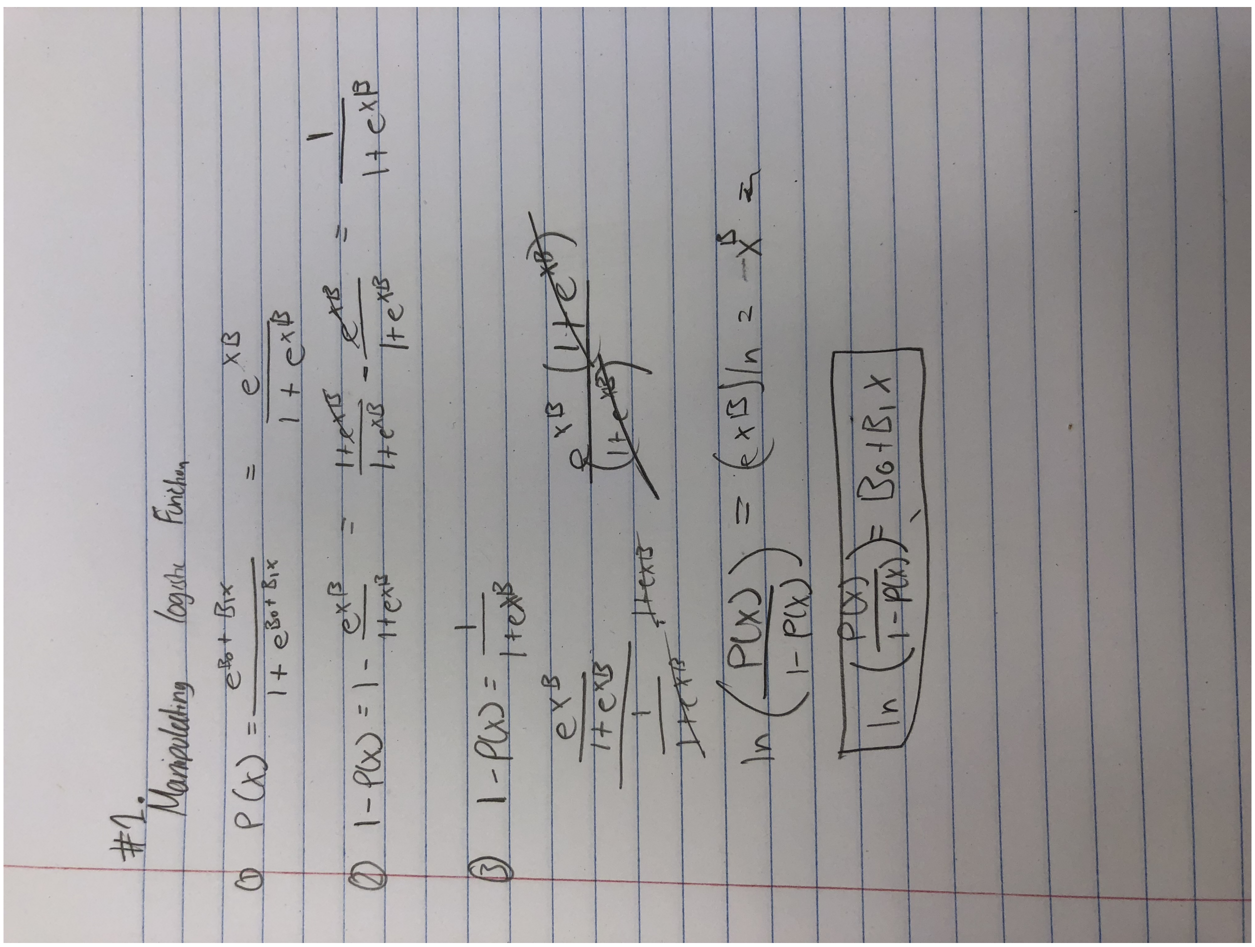
a.

```r
library('ISLR')

library('corrplot')
```

```
## corrplot 0.84 loaded
```

```r
library('ggplot2')
```



#1. Manipulating logistic Function

① $P(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{e^{x\beta}}{1 + e^{x\beta}}$

② $1 - P(x) = 1 - \frac{e^{x\beta}}{1 + e^{x\beta}} = \frac{1 + e^{x\beta}}{1 + e^{x\beta}} - \frac{e^{x\beta}}{1 + e^{x\beta}} = \frac{1}{1 + e^{x\beta}}$

③ $1 - P(x) = \frac{1}{1 + e^{x\beta}}$

$\frac{\frac{e^{x\beta}}{1+e^{x\beta}}}{\frac{1}{1+e^{x\beta}}} = e^{x\beta} \frac{\cancel{(1+e^{x\beta})}}{\cancel{1+e^{x\beta}}}$

$\ln\left(\frac{P(x)}{1 - P(x)}\right) = (e^{x\beta}) \ln e = x\beta$

$\boxed{\ln\left(\frac{P(x)}{1 - P(x)}\right) = \beta_0 + \beta_1 x}$

#b.

```r
data(Weekly)
```

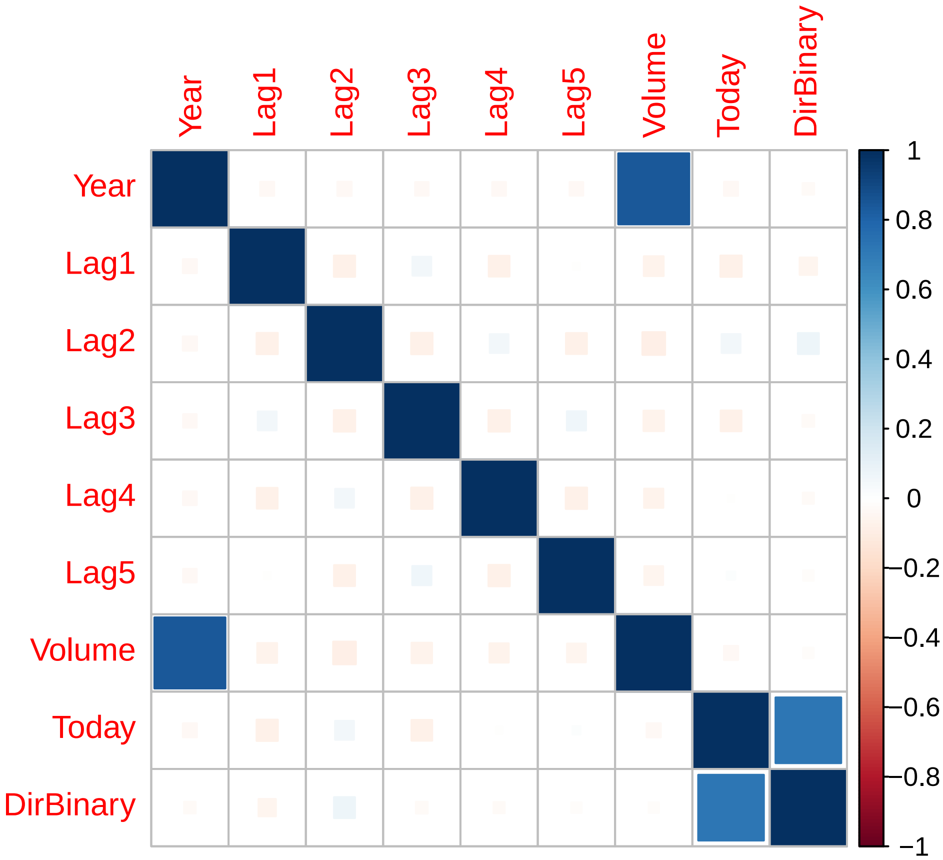
c.

```r
Weekly$DirBinary <- ifelse(Weekly$Direction == "Up",1,0)

corWeekly <- cor(Weekly[,-c(9)])
corWeekly
```

```
##                  Year         Lag1        Lag2        Lag3         Lag4
## Year       1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1      -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2      -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3      -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4      -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5      -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume     0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today     -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
## DirBinary -0.02220025 -0.050003804  0.07269634 -0.02291281 -0.020549456
##                   Lag5      Volume         Today   DirBinary
## Year      -0.030519101  0.84194162 -0.032459894 -0.02220025
## Lag1      -0.008183096 -0.06495131 -0.075031842 -0.05000380
## Lag2      -0.072499482 -0.08551314  0.059166717  0.07269634
## Lag3       0.060657175 -0.06928771 -0.071243639 -0.02291281
## Lag4      -0.075675027 -0.06107462 -0.007825873 -0.02054946
## Lag5       1.000000000 -0.05851741  0.011012698 -0.01816827
## Volume    -0.058517414  1.00000000 -0.033077783 -0.01799521
## Today      0.011012698 -0.03307778  1.000000000  0.72002470
## DirBinary -0.018168272 -0.01799521  0.720024704  1.00000000
```

```r
corrplot(corWeekly,method = "square")
```



d.

The strongest correlation with direction with collinear binary is Today and DirBinary. It has a value of 0.72

e.

```r
library('doBy')
doBy::summaryBy(Year + Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume ~ Direction, data = Weekly, fun = c(mea
```

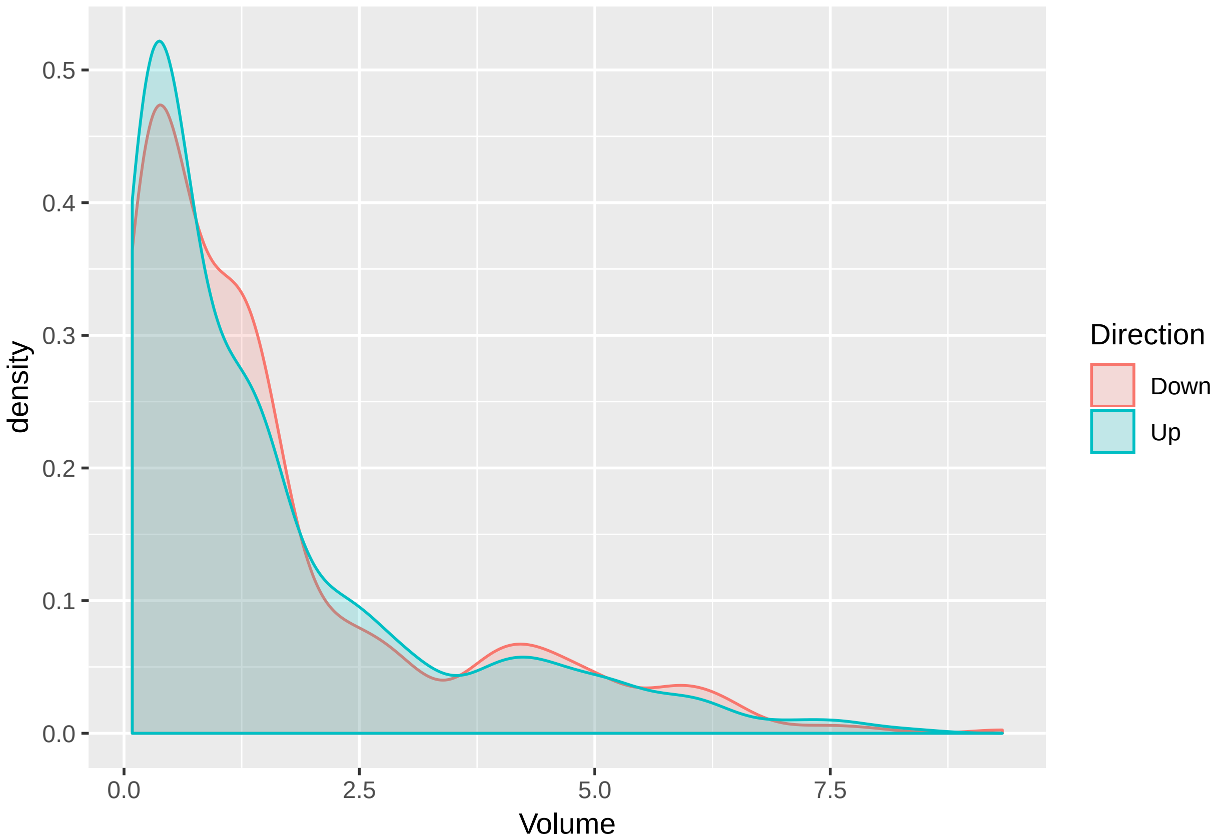
```
##   Direction Year.FUN1  Lag1.FUN1   Lag2.FUN1  Lag3.FUN1 Lag4.FUN1
## 1      Down  2000.198 0.28229545 -0.04042355 0.20764669 0.2000207
## 2        Up  1999.929 0.04521653  0.30428099 0.09885124 0.1024562
##   Lag5.FUN1 Volume.FUN1
## 1 0.1878347    1.608536
## 2 0.1015388    1.547483
```

f.

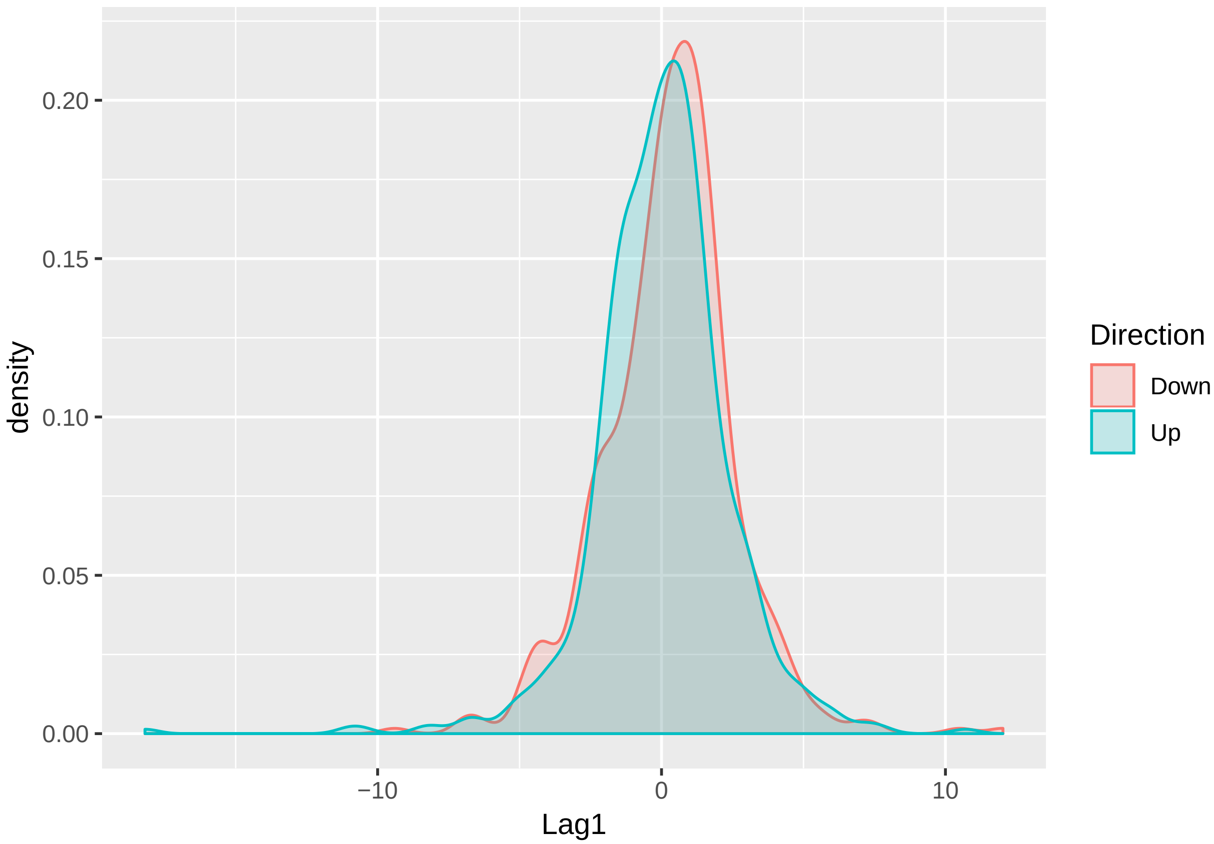
Lag2, Lag1, Lag3, Lag4, Lag5 all have different means

g.

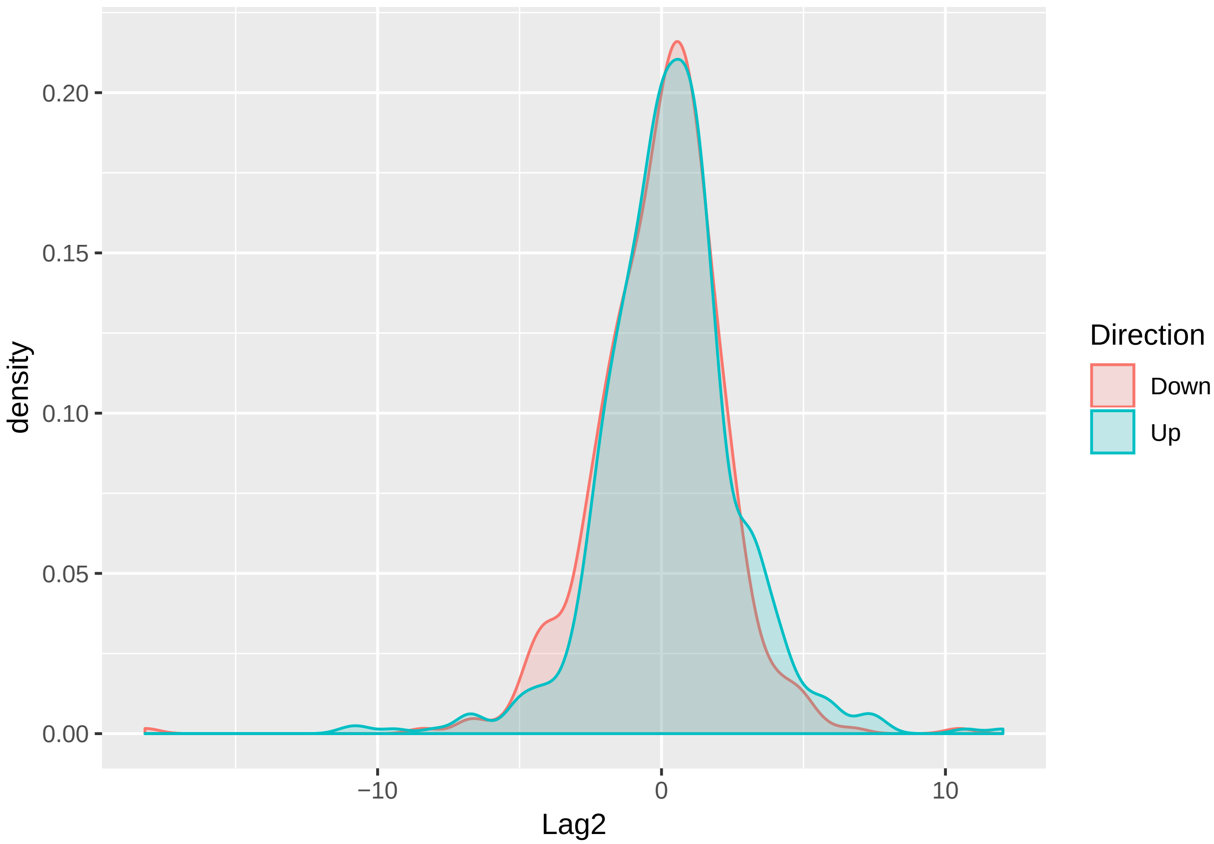
```r
ggplot(Weekly, aes(Volume, fill = Direction, color = Direction)) + geom_density(alpha = 0.2)
```



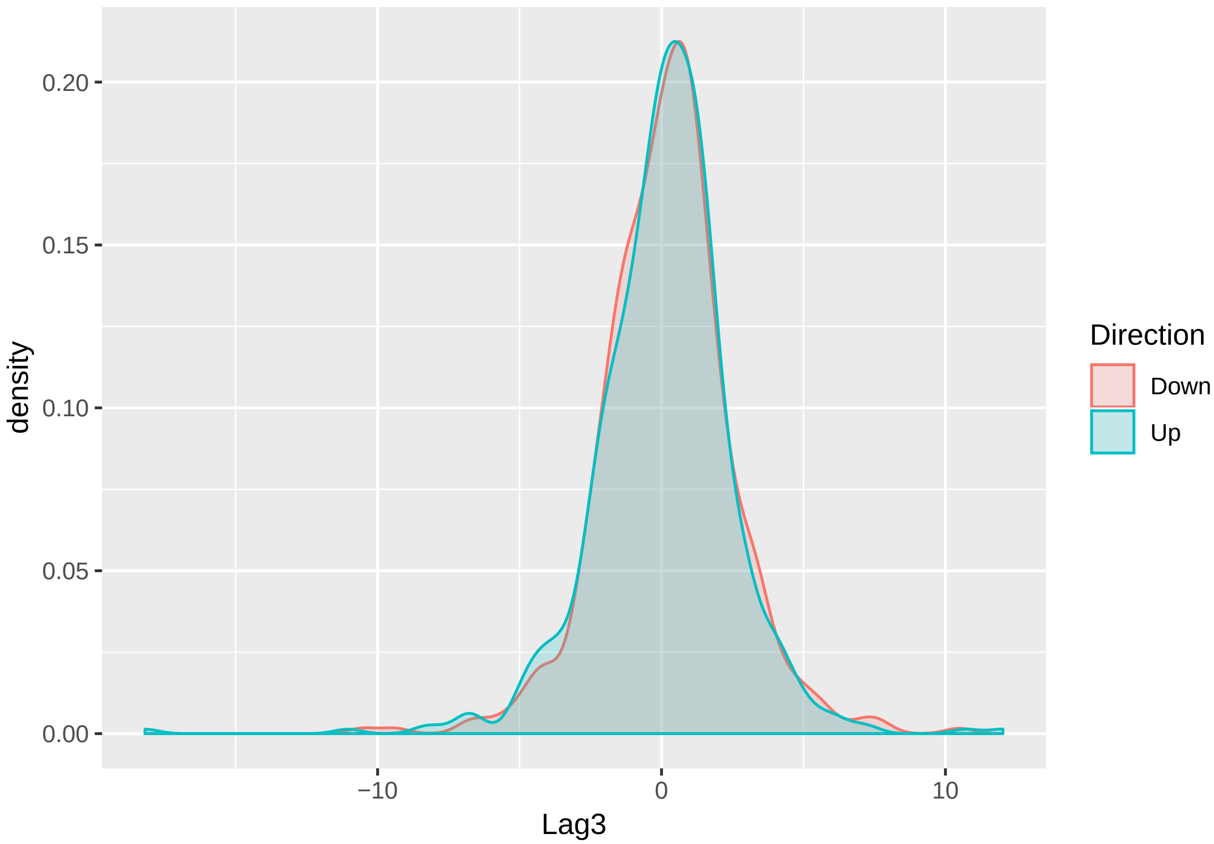

```r
ggplot(Weekly, aes(Lag1, fill = Direction, color = Direction)) + geom_density(alpha = 0.2)
```



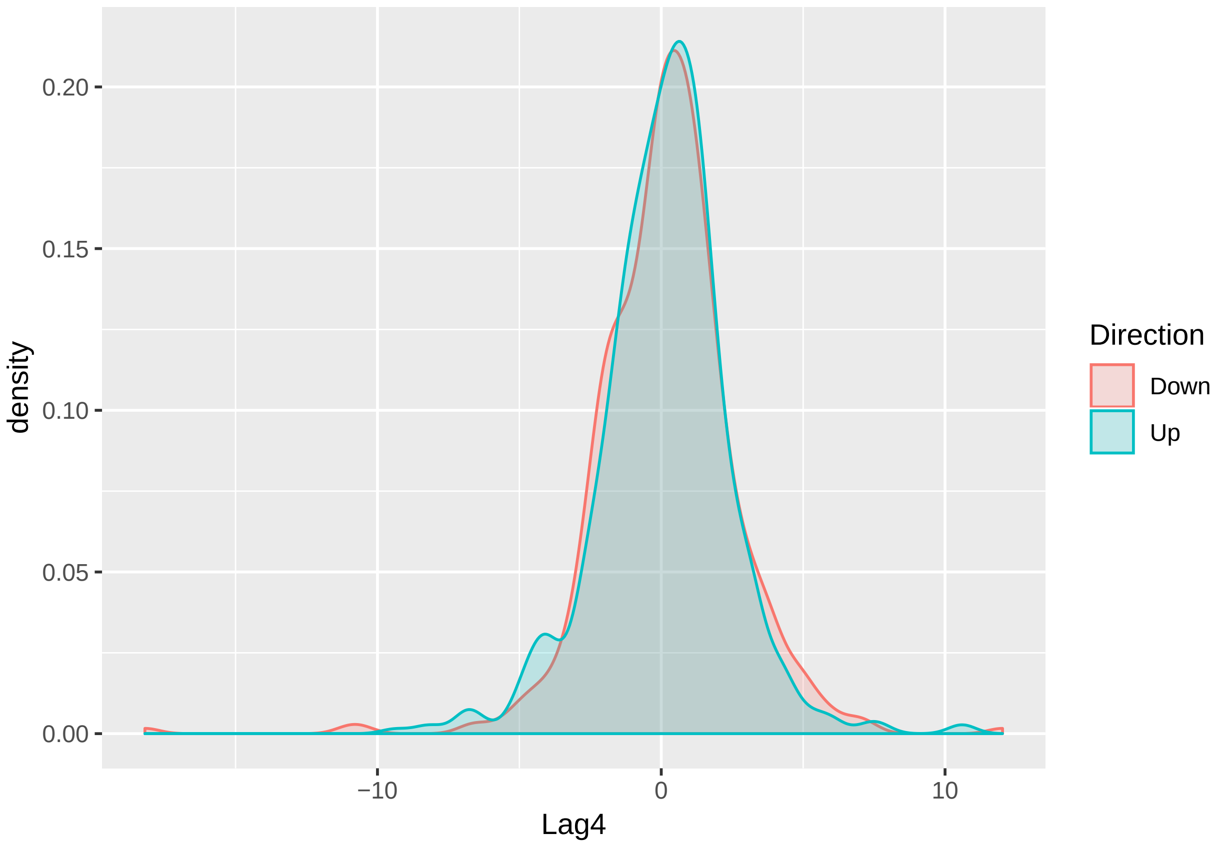

```r
ggplot(Weekly, aes(Lag2, fill = Direction, color = Direction)) + geom_density(alpha = 0.2)
```




```r
ggplot(Weekly, aes(Lag3, fill = Direction, color = Direction)) + geom_density(alpha = 0.2)
```




```
ggplot(Weekly, aes(Lag4, fill = Direction, color = Direction)) + geom_density(alpha = 0.2)
```



3.

a.

```
glmmod <-glm(Direction ~ Lag1+ Lag2 + Lag3 + Lag4 + Lag5, data = Weekly,family = binomial)
```

b.

```
summary(glmmod)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5, family = binomial,
##     data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7297  -1.2574   0.9939   1.0868   1.4671
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.23029    0.06203   3.712 0.000205 ***
```

```
## Lag1        -0.04010    0.02635  -1.522 0.128125
## Lag2         0.06015    0.02674   2.249 0.024503 *
## Lag3        -0.01508    0.02664  -0.566 0.571381
## Lag4        -0.02677    0.02643  -1.013 0.311082
## Lag5        -0.01349    0.02636  -0.512 0.608894
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.7  on 1083  degrees of freedom
## AIC: 1498.7
##
## Number of Fisher Scoring iterations: 4
```

Lag2 is the only variable statistically significant.

c.

```
summary(glmmod)$coef[,4]
```

```
##  (Intercept)         Lag1         Lag2         Lag3         Lag4
## 0.0002053613 0.1281246154 0.0245025891 0.5713805025 0.3110818396
##         Lag5
## 0.6088935425
```

```
coef(glmmod)
```

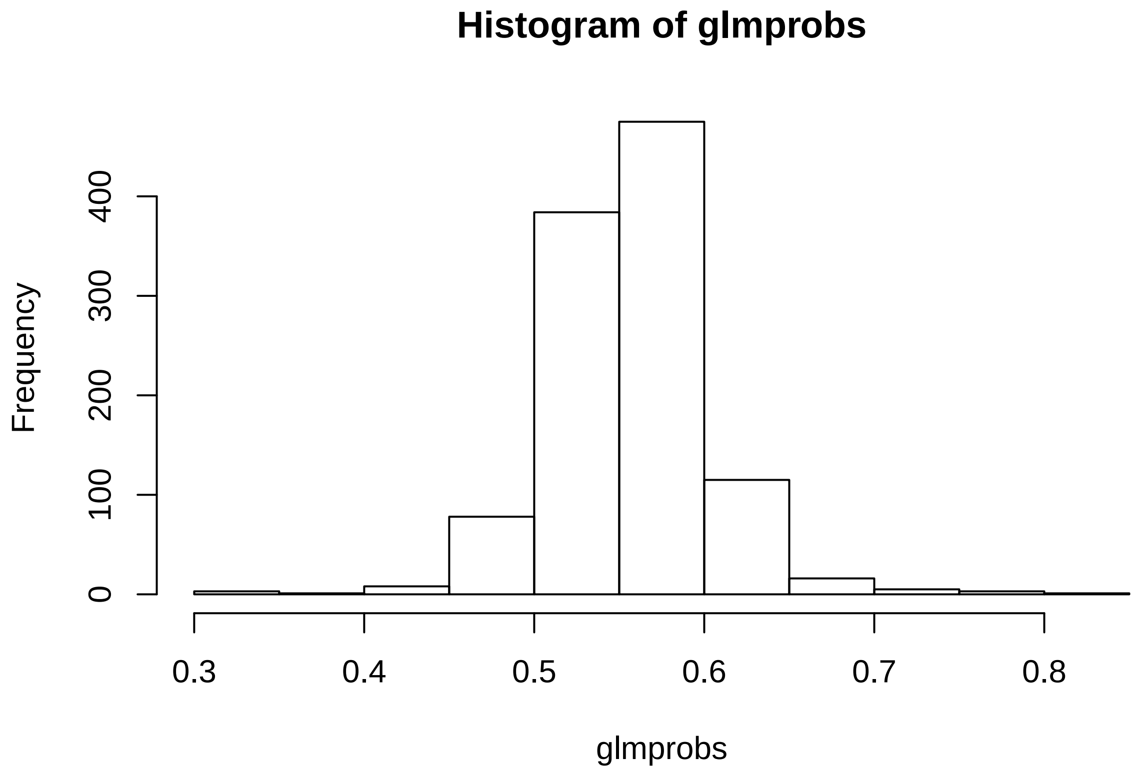
```
## (Intercept)        Lag1        Lag2        Lag3        Lag4        Lag5
##  0.23029037 -0.04009730  0.06015073 -0.01508114 -0.02677052 -0.01348731
```

d.

```
glmprobs <- predict.glm(glmmod, type = "response")
```

e.

```r
hist(glmprobs)
```



Most of them lie between 0.5 and 0.6. Each bar has a respective frequency between 370 to 470.

f.

```r
glmpreds <- rep("Down", nrow(Weekly))
glmpreds[glmprobs > 0.5] <- "Up"
table(glmpreds)
```

```
## glmpreds
## Down   Up
##   90  999
```

g.

```r
table(glmpreds, Weekly$Direction)
```

```
##
## glmpreds Down  Up
##     Down   49  41
##     Up    435 564
```

True positive: 564/605 = 93.2% False negative: 41/605 = 6.5% True negative: 49/484 = 10.12% False positives: 435/485 = 89.9%

We are better at predicting when the market is going to go up, as is indicated by our predicted true positive. We are not as good at predicting true negatives, as we only predicted 10.12%. Our data is more sensitive than specific. Our sensitive rate is 93.2% and our specific rate is 6.8%, meaning we are better at predicting when it will go up as opposed to down.