Q30-WilliamKennedy-300015367

William Kennedy

2023-11-11

Consider the Weekly data set. It contains 1,089 weekly stock market returns for 21 years, from the beginning of 1990 to the end of 2010.

1.Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

```
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
```

```
library("ggplot2")
Weekly = read.csv("Weekly.csv")
Weekly$Direction = ifelse(Weekly$Direction == "Up",1,0)

summary(Weekly)
```

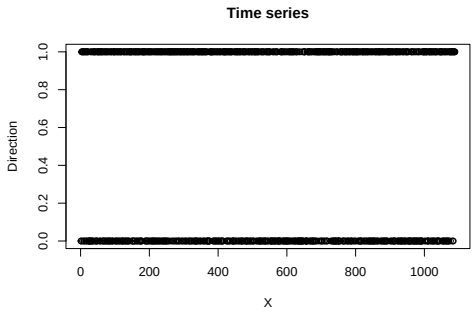
```
##        X             Year           Lag1               Lag2
##  Min.   :   1   Min.   :1990   Min.   :-18.1950   Min.   :-18.1950
##  1st Qu.: 273   1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540
##  Median : 545   Median :2000   Median :  0.2410   Median :  0.2410
##  Mean   : 545   Mean   :2000   Mean   :  0.1506   Mean   :  0.1511
##  3rd Qu.: 817   3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090
##  Max.   :1089   Max.   :2010   Max.   : 12.0260   Max.   : 12.0260
##       Lag3               Lag4               Lag5              Volume
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747
##  1st Qu.: -1.1580   1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202
##  Median :  0.2410   Median :  0.2380   Median :  0.2340   Median :1.00268
##  Mean   :  0.1472   Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462
##  3rd Qu.:  1.4090   3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373
##  Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821
##      Today            Direction
##  Min.   :-18.1950   Min.   :0.0000
```

```
##  1st Qu.: -1.1540   1st Qu.:0.0000
##  Median :  0.2410   Median :1.0000
##  Mean   :  0.1499   Mean   :0.5556
##  3rd Qu.:  1.4050   3rd Qu.:1.0000
##  Max.   : 12.0260   Max.   :1.0000
```

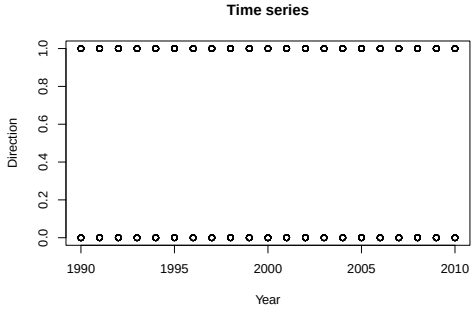
```
for(i in 1:(ncol(Weekly)-1)) {
  col = Weekly[,i]
  print(paste("Mean of",colnames(Weekly)[i],"is",mean(col)))
  print(paste("Median of",colnames(Weekly)[i],"is",median(col)))
  print(paste("Variance of",colnames(Weekly)[i],"is",var(col)))
  print("")

  plot(col,Weekly$Direction,xlab=colnames(Weekly)[i],ylab="Direction",main="Time series")
}
```

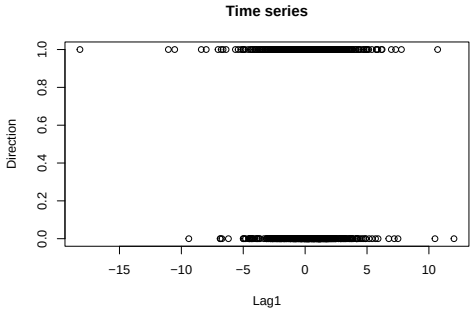
```
## [1] "Mean of X is 545"
## [1] "Median of X is 545"
## [1] "Variance of X is 98917.5"
## [1] ""
```



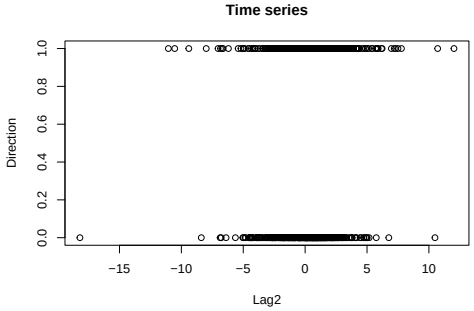

```
## [1] "Mean of Year is 2000.04866850321"
## [1] "Median of Year is 2000"
## [1] "Variance of Year is 36.3992836115162"
## [1] ""
```



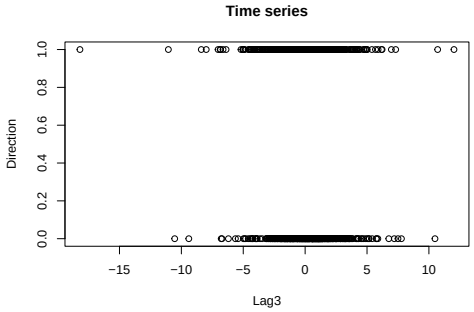

```
## [1] "Mean of Lag1 is 0.150584940312213"
## [1] "Median of Lag1 is 0.241"
## [1] "Variance of Lag1 is 5.555508029773"
## [1] ""
```



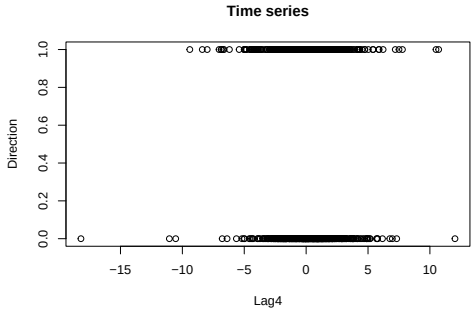

```
## [1] "Mean of Lag2 is 0.151078971533517"
## [1] "Median of Lag2 is 0.241"
## [1] "Variance of Lag2 is 5.55664749008129"
## [1] ""
```



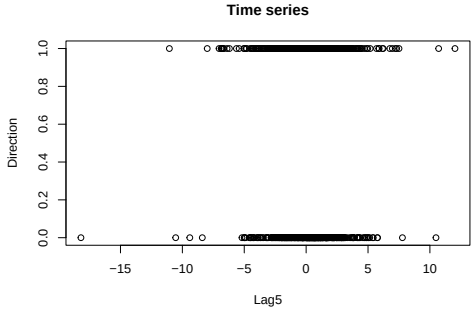

```
## [1] "Mean of Lag3 is 0.147204775022957"
## [1] "Median of Lag3 is 0.241"
## [1] "Variance of Lag3 is 5.57196960968306"
## [1] ""
```



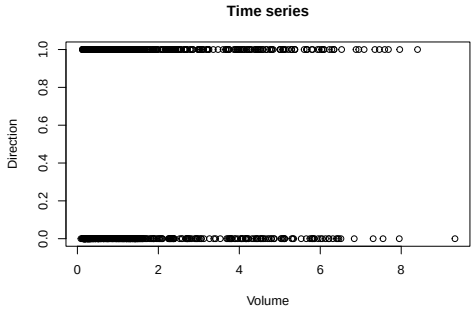

```
## [1] "Mean of Lag4 is 0.145818181818182"
## [1] "Median of Lag4 is 0.238"
## [1] "Variance of Lag4 is 5.57091625"
## [1] ""
```



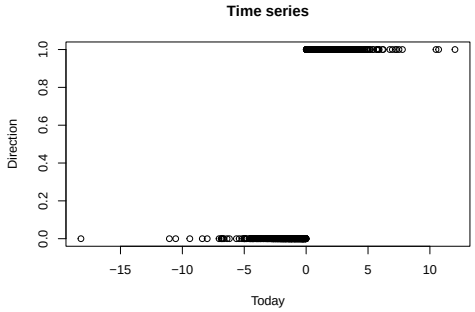

```
## [1] "Mean of Lag5 is 0.139892561983471"
## [1] "Median of Lag5 is 0.234"
## [1] "Variance of Lag5 is 5.5756653184097"
## [1] ""
```




```
## [1] "Mean of Volume is 1.57461762552801"
## [1] "Median of Volume is 1.00268"
## [1] "Variance of Volume is 2.84474238640386"
## [1] ""
```




```
## [1] "Mean of Today is 0.14989898989899"
## [1] "Median of Today is 0.241"
## [1] "Variance of Today is 5.55510671221405"
## [1] ""
```




```
Dir = Weekly$Direction
Lag1 = Weekly$Lag1
Lag2 = Weekly$Lag2
Lag3 = Weekly$Lag3
Lag4 = Weekly$Lag4
Lag5 = Weekly$Lag5
Volume = Weekly$Volume

log.reg = glm(Dir ~ Lag1+Lag2+Lag3+Lag4+Lag5+Volume,data=Weekly, family = binomial)

summary(log.reg)
```

```
##
## Call:
## glm(formula = Dir ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
##     family = binomial, data = Weekly)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## 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: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

Lag2, the previous 2 week return, has the heaviest coefficient weight in determing the direction and based on it's p-value is statistically significant.

3. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

```
prob = predict(log.reg , type = "response")
pred = rep ("Down", 1089)
pred[prob > .5] = "Up"
table(pred , Weekly$Direction)
```

```
##
## pred     0   1
##   Down  54  48
##   Up   430 557
```

The confusion matrix of the logistic regression model tells us the model made 54 true positive predictions, 48 false negative predictions, 430 false positive predictions, and 557 true negative reports

The number of incorrect over correct predictions is $\frac{478}{611} \approx 0.7823$ and the percentage of correct predictions is $\frac{611}{1089} \approx 0.561$.

4.Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

```
train = subset(Weekly,Year<2009)
Smarket.2009.2010 <- Weekly[!train , ]
Direction.2009.2010 <- Smarket.2009.2010$Direction[!train]

log.reg2 = glm(train$Direction~train$Lag2,data=train,family=binomial)

prob2 = predict(log.reg2 , Smarket.2009.2010,type = "response")
```

```
## Warning: 'newdata' had 441 rows but variables found have 985 rows
```

7. Repeat 4. using kNN with $k = 1$.