HW5

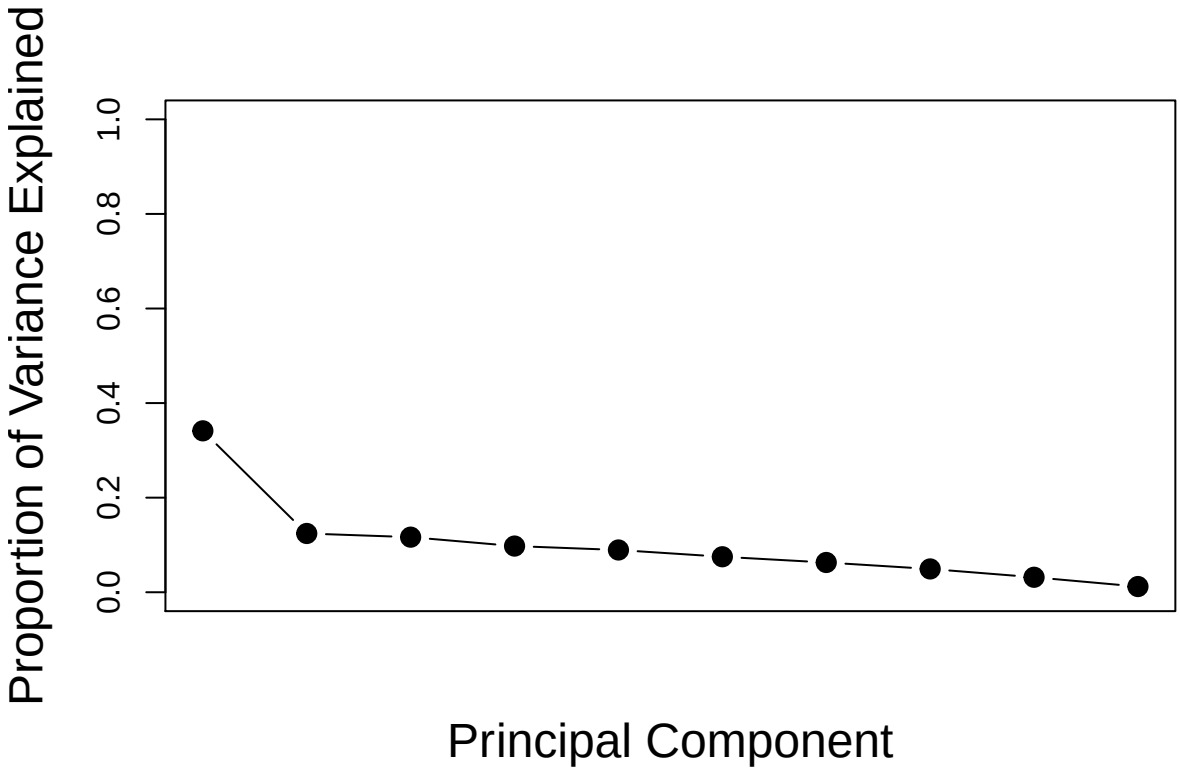
2/10/2022

Q1

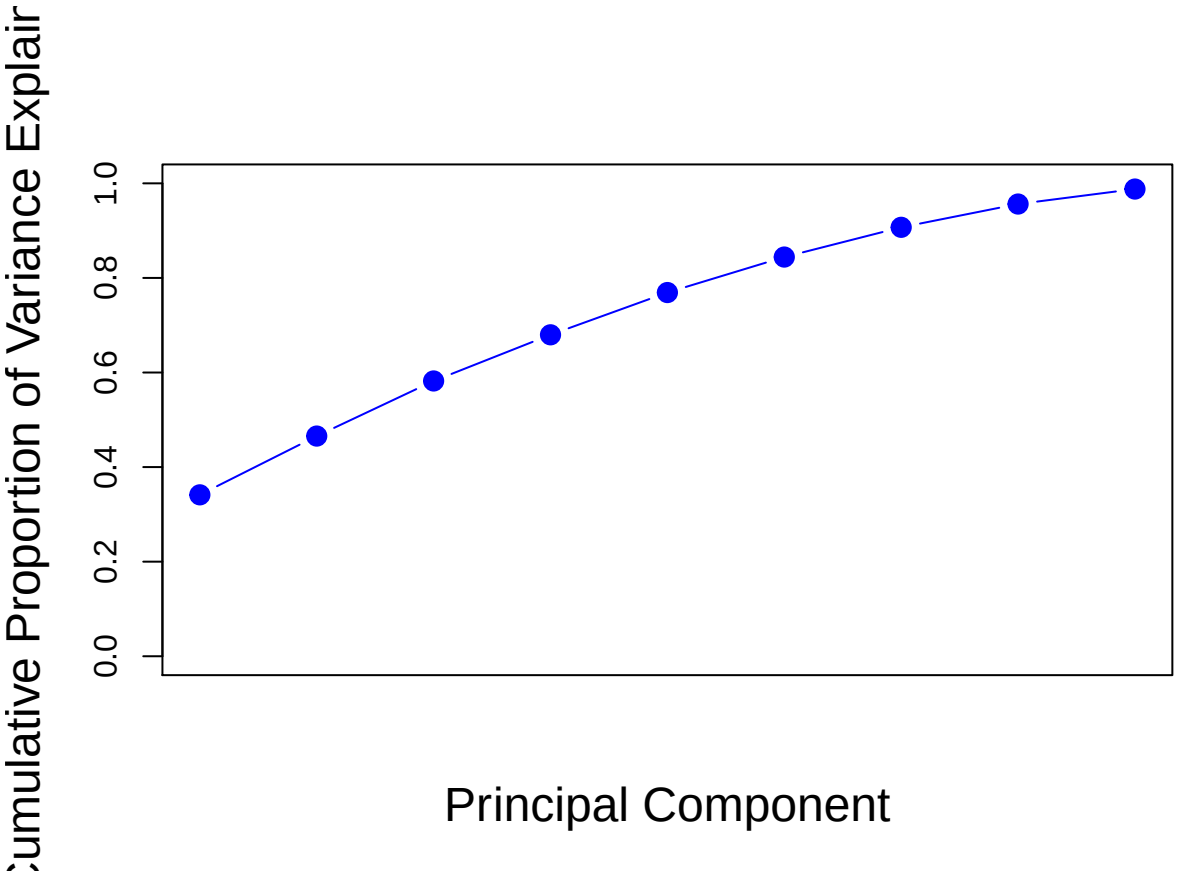
a

```
data <- read.delim("Places_Rated.txt", header =FALSE, sep = '')
df <- data.frame(data)
names(df) <- c("Climate and Terrain", "Housing",
               "Health Care & the Environment", "Crime",
               "Transportation", "Education", "The Arts",
               "Recreation", "Economics", "index")
standard_df <- scale(df)
standard_cov <- cov(standard_df)
standard_ev <- eigen(standard_cov)
standard_prop <- c()
for ( i in 1:10) {
  prop <- (standard_ev$values[i]) / (sum(standard_ev$values))
  standard_prop <- append(standard_prop, prop)
}
standard_cprop <- c()
cprop <- 0
for ( i in 1:9) {
  cprop <- cprop + standard_prop[i]
  standard_cprop <- append(standard_cprop, cprop)
}

plot(standard_prop, xlab="Principal Component",
     ylab="Proportion of Variance Explained",
     ylim=c(0,1), xaxt="n", type='b', cex=2, pch=20, cex.lab=1.5)
```

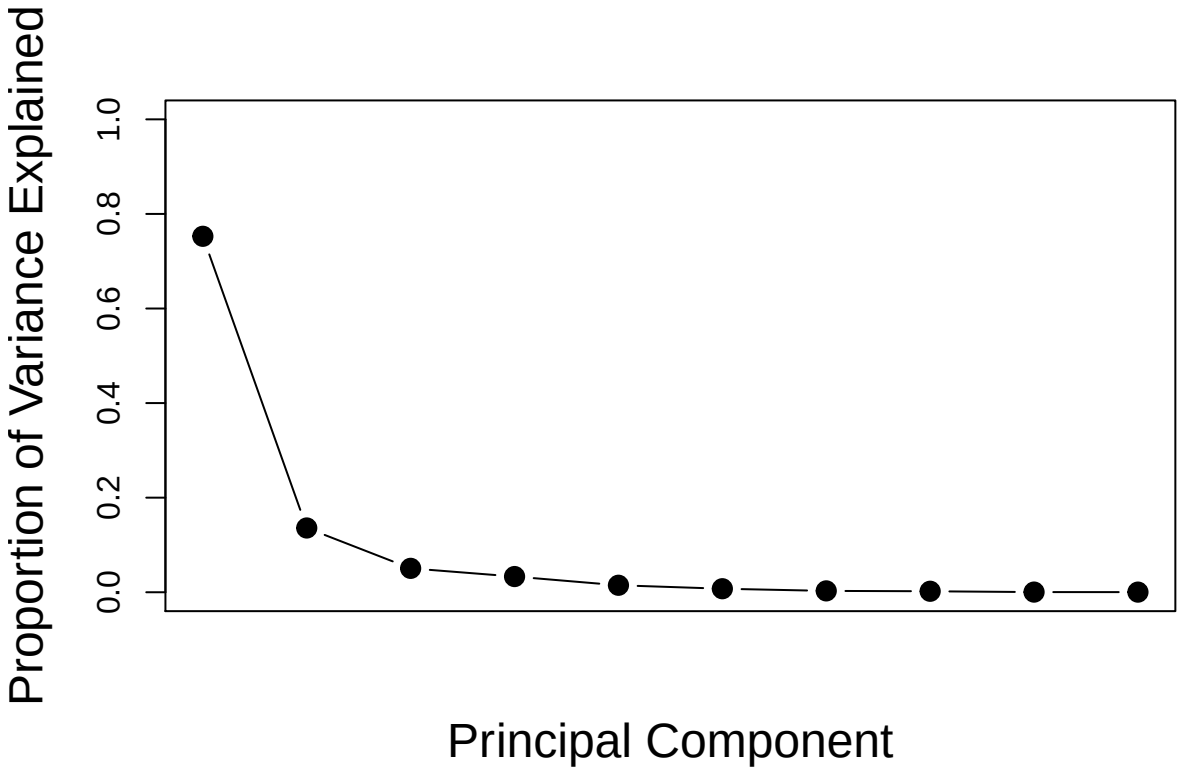



```
plot(standard_cprop, xlab="Principal Component",
     ylab="Cumulative Proportion of Variance Explained",
     ylim=c(0,1), xaxt="n", type="b", col="blue", cex=2,
     pch=20, cex.lab=1.5)
```

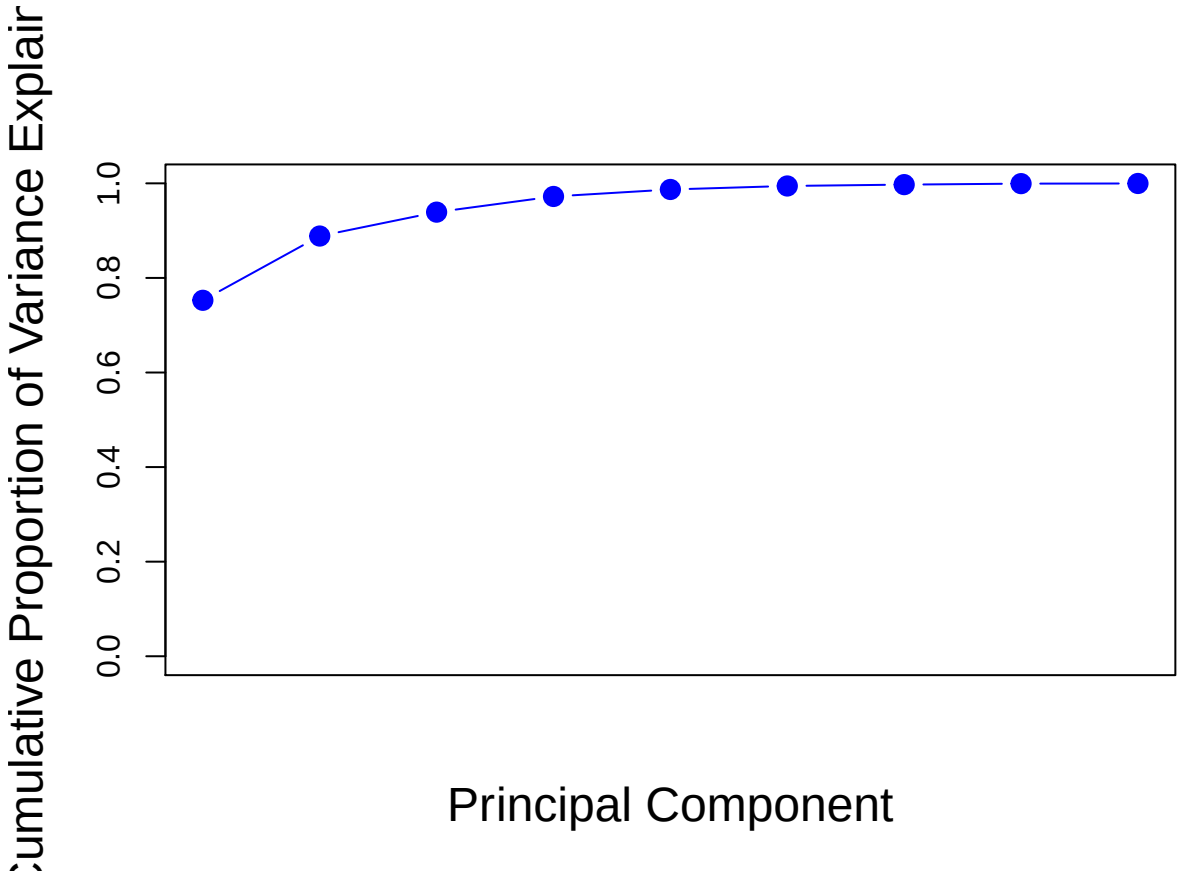



```
raw_cov <- cov(df)
raw_ev <- eigen(raw_cov)
raw_prop <- c()
for ( i in 1:10) {
  prop <- (raw_ev$values[i]) / (sum(raw_ev$values))
  raw_prop <- append(raw_prop, prop)
}
raw_cprop <- cumsum(raw_prop[1:9])

plot(raw_prop, xlab="Principal Component",
     ylab="Proportion of Variance Explained", ylim=c(0,1),
     xaxt="n", type='b', cex=2, pch=20, cex.lab=1.5)
```




```
plot(raw_cprop, xlab="Principal Component",
     ylab="Cumulative Proportion of Variance Explained",
     ylim=c(0,1), xaxt="n", type="b", col="blue", cex=2,
     pch=20, cex.lab=1.5)
```



b

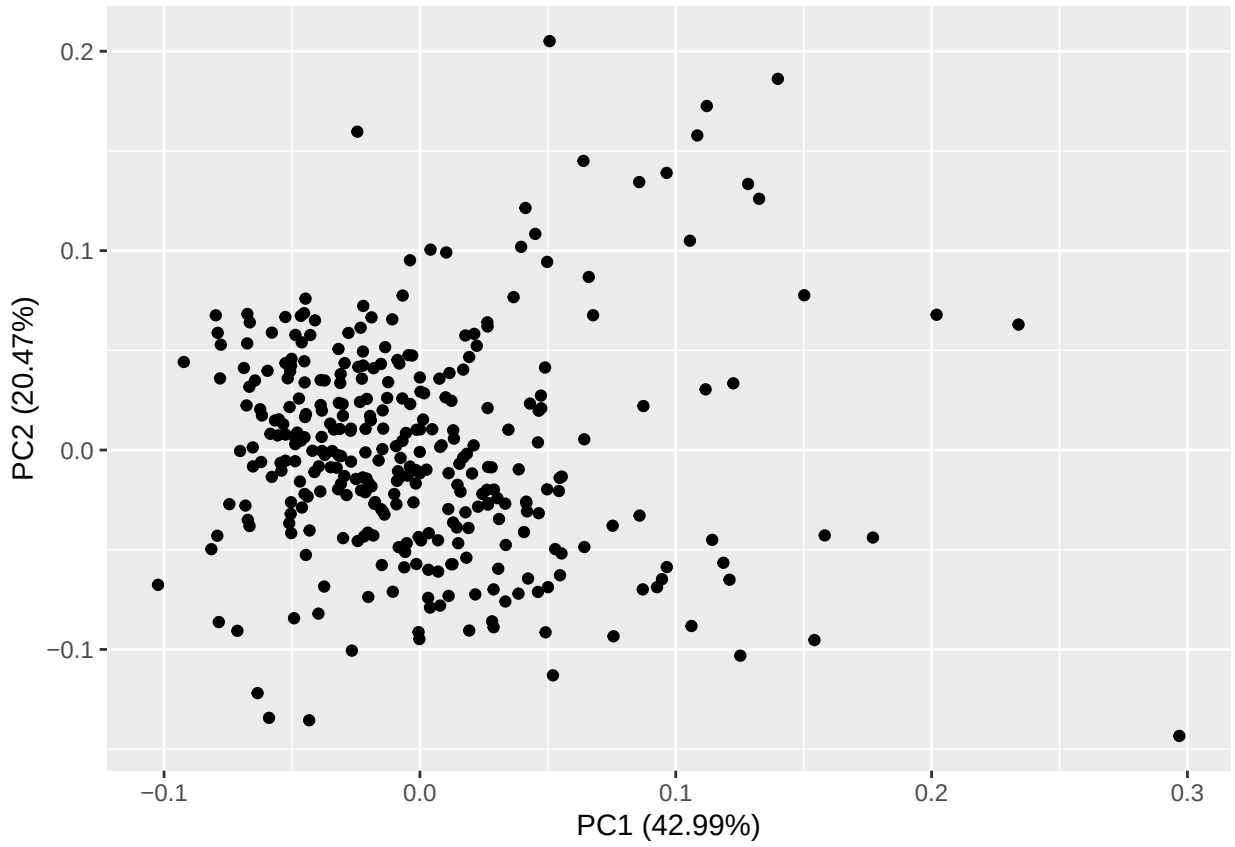
```
library("ggplot2")
library("ggfortify")
```

```
## Warning: package 'ggfortify' was built under R version 4.0.5
```

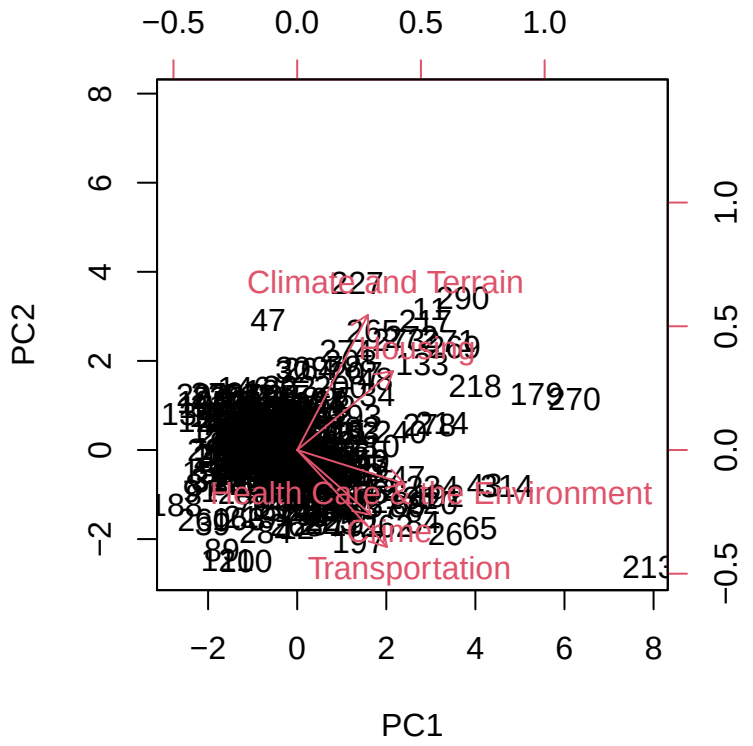
```
pca_result <- prcomp(df[,1:5], scale=TRUE)
pca_result$rotation
```

```
##                                     PC1        PC2        PC3        PC4
## Climate and Terrain           0.3568742  0.6823118 -0.3438239  0.4810650
## Housing                       0.4842331  0.3983105  0.3493474 -0.3856060
## Health Care & the Environment 0.5415405 -0.1704116  0.2556157 -0.3307141
## Crime                         0.3729212 -0.3265431 -0.7959444 -0.2794155
## Transportation                0.4536769 -0.4900294  0.2467278  0.6576015
##                                     PC5
## Climate and Terrain            0.2396936
## Housing                       -0.5797684
## Health Care & the Environment  0.7092127
## Crime                         -0.2066368
## Transportation                -0.2464430
```

```
autoplot(pca_result, data=df, color='black')
```




```
biplot(pca_result, scale=0)
```



Question #2

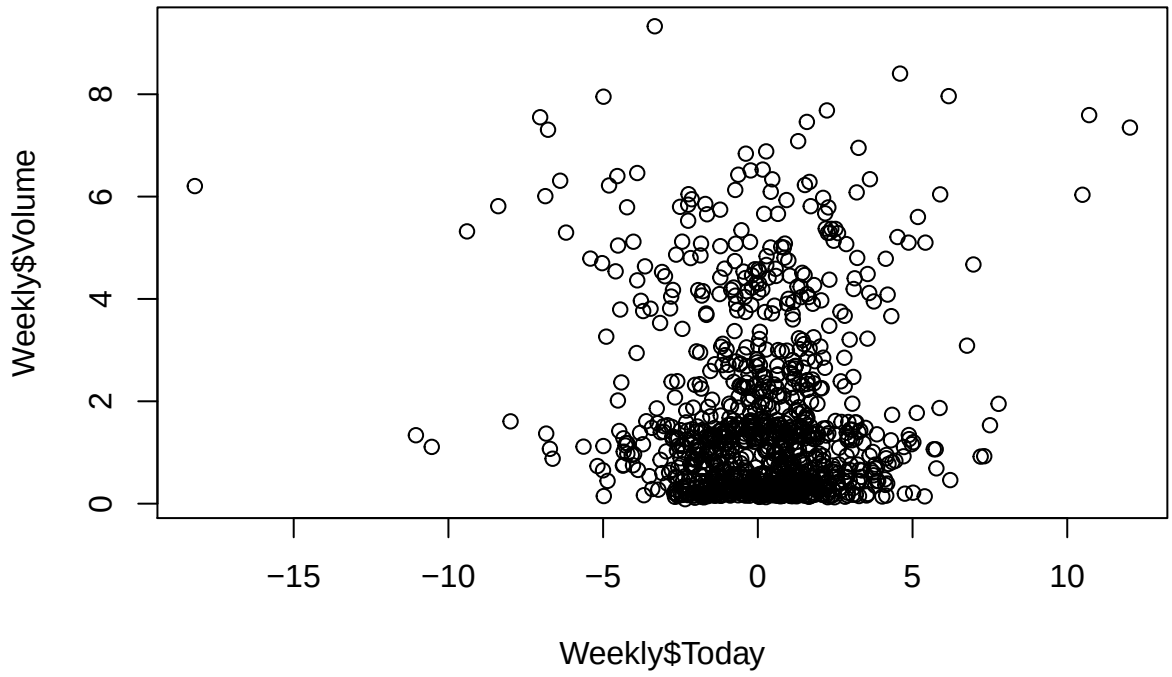
```
summary(Weekly)
```

```
##       Year           Lag1                Lag2                Lag3          
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472  
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090  
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260  
##       Lag4                Lag5              Volume            Today          
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540  
##  Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260  
##  Direction
##  Down:484
##  Up  :605
##
##
```

```
##
##
```

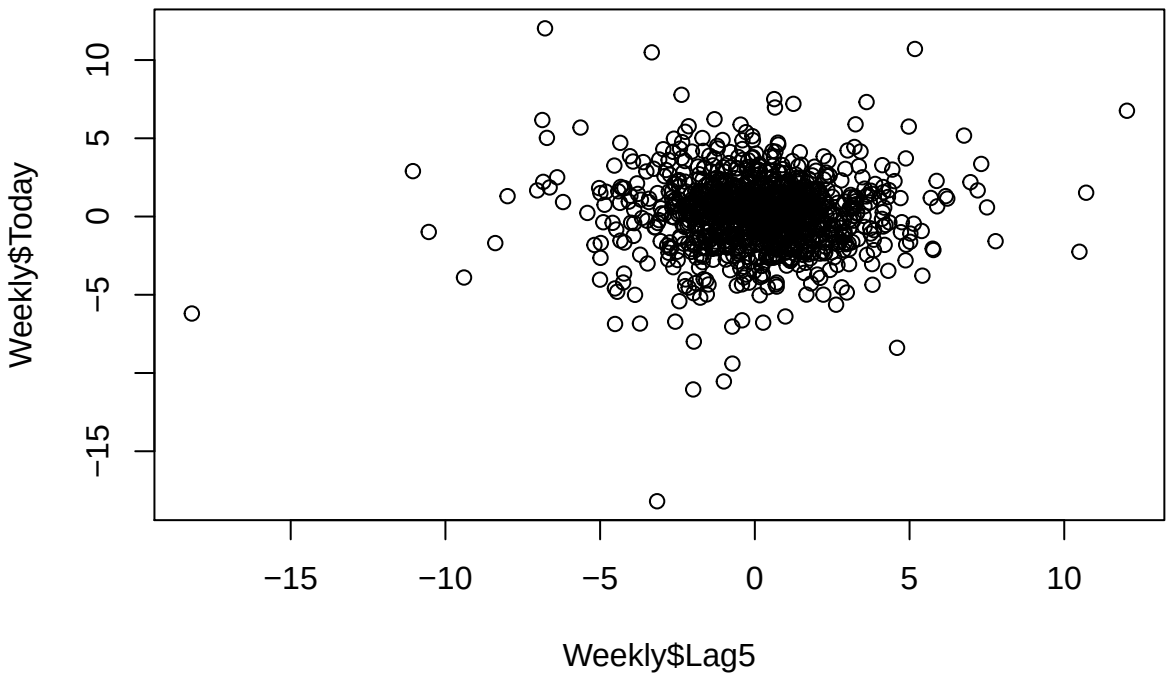
From the numerical summary, we can see that all of the legs have a positive median and mean, signifying on average the stock market grows in a positive direction. This is also evident from the direction variable, with there being significantly more weeks where there was a net positive change.

```
plot(Weekly$Today, Weekly$Volume)
```



From the plot we can see that that if there is a significant change in the price during the week, it would most likely be neagtive. Also, weeks with larger volumes have larger changes in price.

```
plot(Weekly$Lag5, Weekly$Today)
```



From the graph there is minimal correlation between the past 5 weeks of the market and how the market will do this week.

Question 2b)

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

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## 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
```

None of the variables seem statistically significant, but Lag2 seems to be the most useful out of all of them.

Question 2c

```
Probs = predict(glm.fit, type='response')
contrasts(Weekly$Direction)
```

```
##      Up
## Down  0
## Up    1
```

```
Pred_trend = ifelse(Probs>0.5, "Up", "Down")
table(Pred_trend, Weekly$Direction)
```

```
##
## Pred_trend Down  Up
##       Down   54  48
##       Up    430 557
```

```
mean(Pred_trend == Weekly$Direction)
```

```
## [1] 0.5610652
```

Fraction of correct predict is 0.561. This model is predicting to much "up" since the dataset is unbalanced.

Question 2d

```
Data_train = Weekly[Weekly$Year <= 2008,]
Data_test = Weekly[Weekly$Year > 2008,]
glm_fits2 = glm(Direction ~ Lag2, data = Data_train, family = binomial)
Probs_2 = predict(glm_fits2, Data_test, type = "response")
Pred_trend2 = ifelse(Probs_2>0.5, "Up", "Down")
table(Pred_trend2, Data_test$Direction)
```

```
##
## Pred_trend2 Down Up
##        Down    9  5
##        Up     34 56
```

```
mean(Pred_trend2 == Data_test$Direction)
```

```
## [1] 0.625
```