EXECUTIVE SUMMARY

The data in Table 7.6 show the survival times from diagnosis of patients suffering from leukemia and the values of two explanatory variables, the white blood cell count (wbc) and the presence or absence of a morphological characteristic of the white blood cells (ag) (the data are available in package MASS, Venables and Ripley, 2002)

1 .1) Define a binary outcome variable according to whether or not patients lived for at least 24 weeks after diagnosis?

To do this in the first place I'll make sure that I have the necessary packages which is MASS and the data I'm looking for dataset called leuk short for Leukemia. My binary outcome variable to represent whether a patient survived 24 weeks after diagnosis is survivalBinary24Weeks.

wbc	ag	time	survivalBinary24Weeks
2300	present	65	1
750	present	156	1
4300	present	100	1
2600	present	134	1
6000	present	16	0
10500	present	108	1
10000	present	121	1
17000	present	4	0
5400	present	39	1
7000	present	143	1
9400	present	56	1
32000	present	26	1
35000	present	22	0
100000	present	1	0
100000	present	1	0
52000	present	5	0
100000	present	65	1

2 Fit a logistic regression model to the data. It may be advisable to transform the very large white blood counts to avoid regression coefficients very close to 0 (and odds ratios very close to 1), and fit a model that contains only the two explanatory variables may not be adequate for these data.

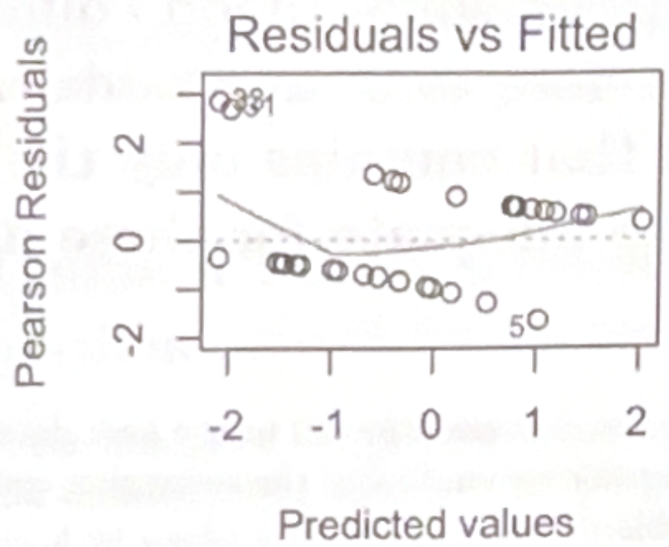
I did a log transformation on the wbc variableTo fit a logistic regression model to the leuk dataset and address the issue of large white blood cell counts. This transformation makes the regression coefficients more interpretable.

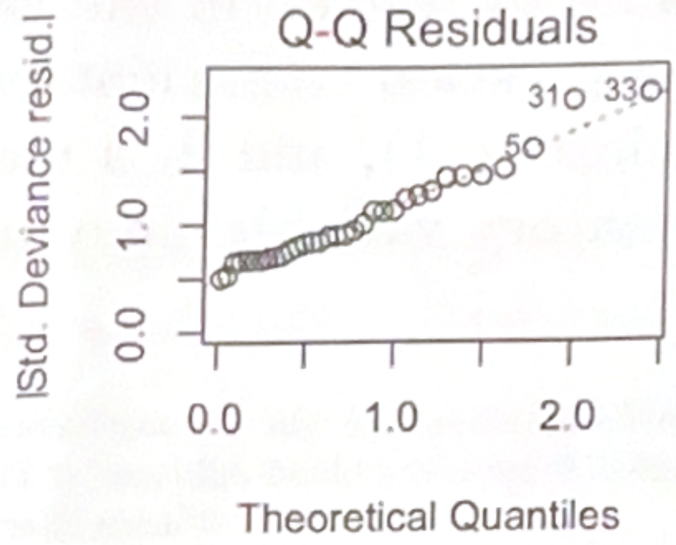
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.4556     2.9821   1.159   0.2466
log_wbc      -0.4822     0.3149  -1.531   0.1257
agpresent     1.7621     0.8093   2.177   0.0295 *
---
```

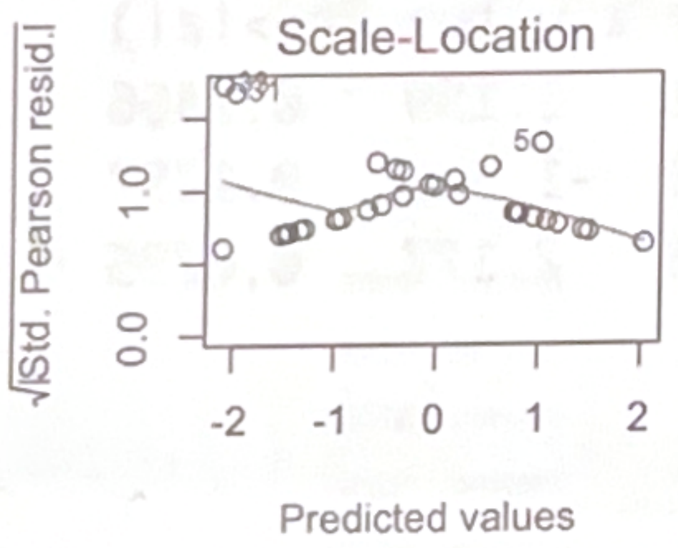
Font?

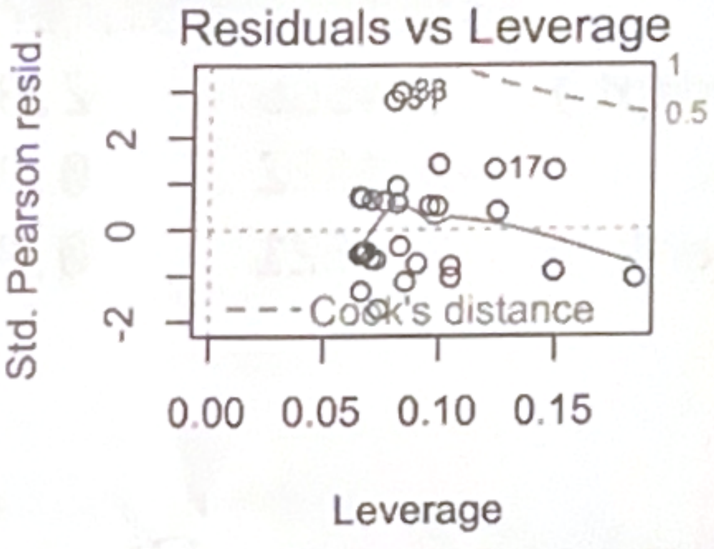
3 Construct some graphics useful in the interpretation of the final model you fit

From what I found there is some non-linearity in the Residuals vs Fitted plot meaning that the relationship between the predictors and the response may not be well captured by the logistic regression model.The Scale-Location plot is pointing to the fact that the variance of residuals is not constant across the fitted values.For me to come up with these visuals the plot function.



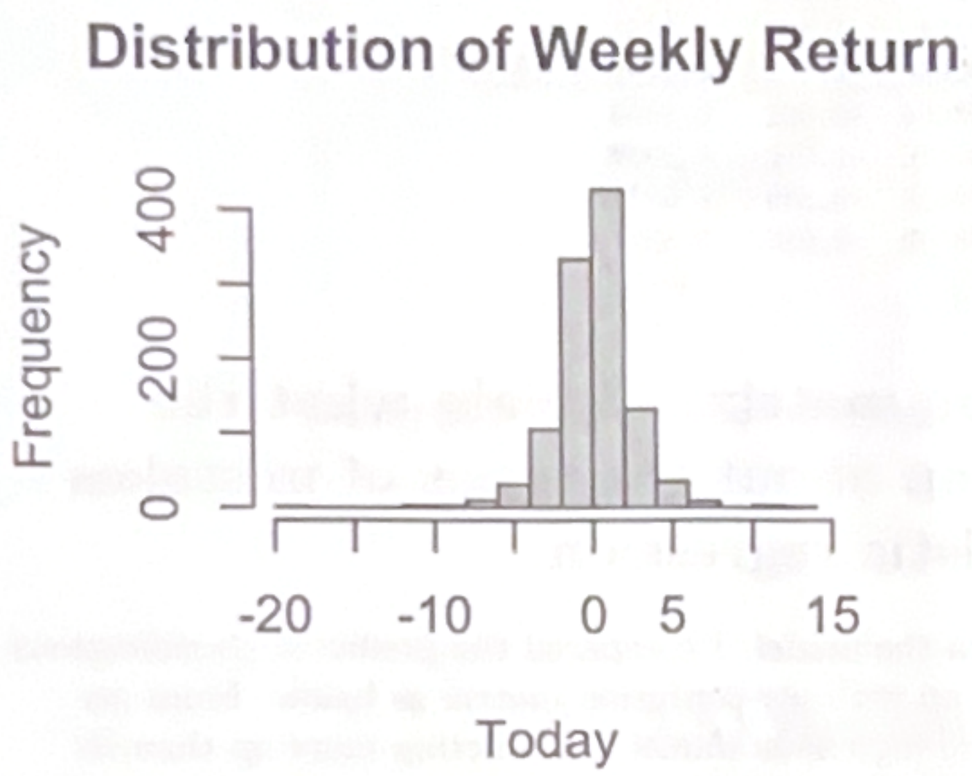


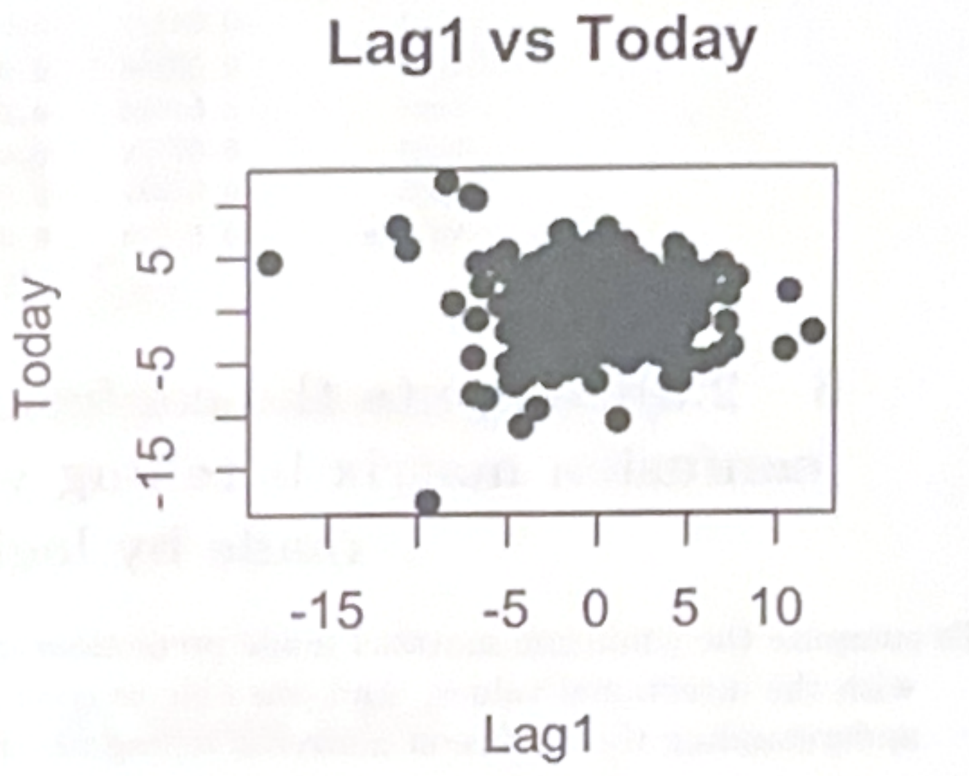


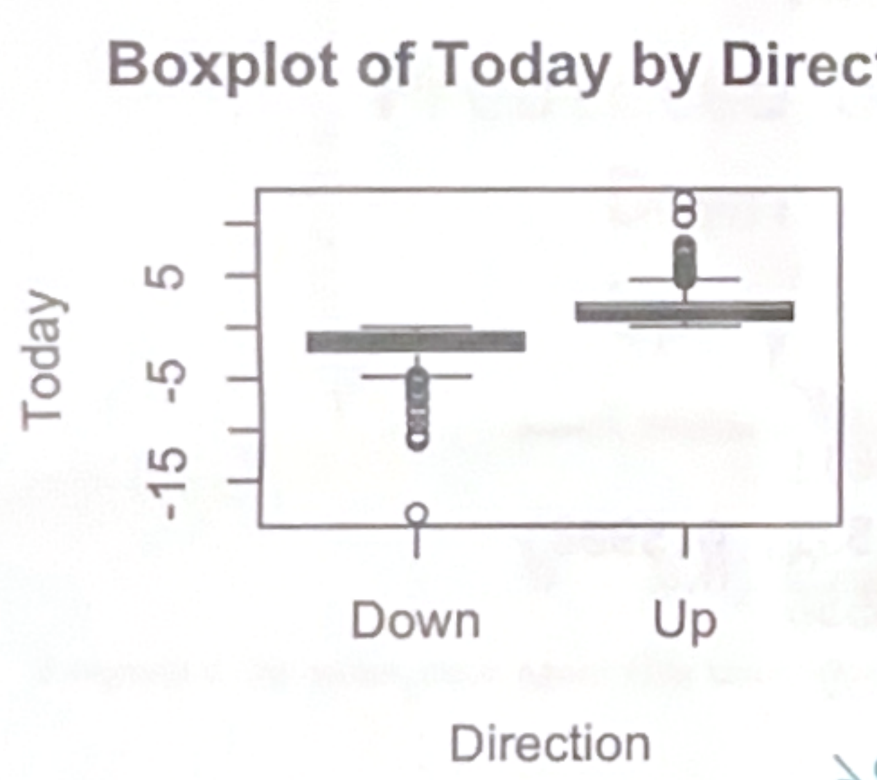


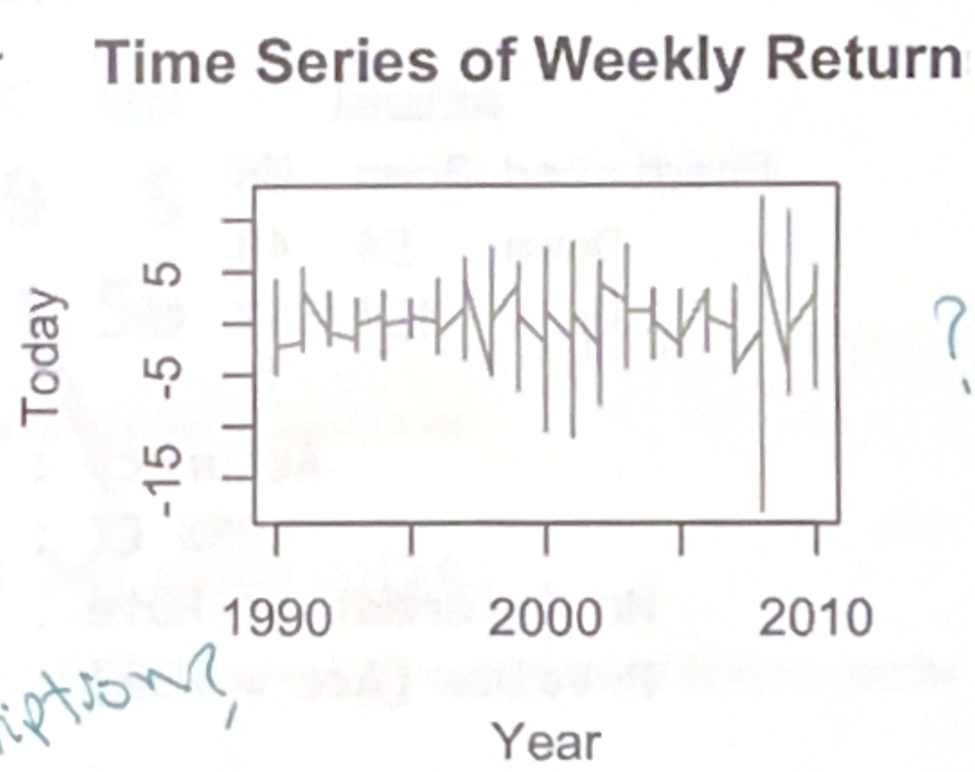
4 2.1.) This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010. Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

I produced some numerical and graphical summaries of the Weekly data.My histogram shows the distribution the today variable which represents the weekly return. Its showing a normal distribution of weekly returns. The scatter plot shows the relationship between Lag1 and Today from my findings there appears to be no relation between the two. The Boxplot is showing me the distribution on today based on direction, then lastly the last graph is a plot showing weekly return from 1990 to 2010. There is no clear improvement or loss as there are continousd fluctuations.









description?

?

5 2.2) Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

To do this I used the glm function then displayed the summary of the model. I made sure to use direction the five lag variables and it came out like this. I don't think any of the predictors appear to be statistically significant, all p values are greater than 0.05 which goes on to say prtedictin the market direction will be difficult.

```
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
```

6 2.3)Compute the confusion matrix. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

To compute the confusion matrix I made predictions on the model. I compared the predicted classifications with the directional values. And was able to come up with my confusion matrix as below. From my understanding the confusion matrix is telling me the regression model is predicting more up than its supposed to be.

```
          Actual
Predicted Down  Up
     Down   54  48
     Up    430 557

               Accuracy : 0.5611
                 95% CI : (0.531, 0.5908)
    No Information Rate : 0.5556
    P-Value [Acc > NIR] : 0.369
```

7 2.4) 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).

To fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor and Compute the confusion matrix and the overall fraction of correct predictions. I firstly split the data into training and testing sets as my training set is from 1990 to 2008 and my testing is from 2009 to 2010. I then went on to fit my model whilst using lag2 as my predictor. Lastly I computed my confusion matric and the fraction of correct predictions.

```
             Actual
Predicted Down Up
     Down    9  5
     Up     34 56
```

```
Overall Fraction of Correct Predictions (Accuracy): 0.625
```

8 2.5) Repeat (d) using LDA

I repeated the same step before this time with Linear Discriminant Analysis and got the same results.

```
             Actual
Predicted Down Up
     Down    9  5
     Up     34 56
```

```
Overall Fraction of Correct Predictions (LDA Accuracy): 0.625
```

9 2.6) Repeat (d) using QDA

I repeated the same steps again this time with Quadratic Discriminant Analysis and got different results

```
             Actual
Predicted Down Up
     Down    0  0
     Up     43 61
>
```

```
Overall Fraction of Correct Predictions (QDA Accuracy): 0.5865385
```

10 2.7) Which of these methods appears to provide the best results on this data?

Based on the results I have gotten Logistic Regression and LDA provide the most accurate results with highest accuracy. QDA's accuracy did not beat that of the logistic regression and the LDA.

11 Appendix

- R code

```
install.packages("MASS")
library(MASS)
data("leuk", package = "MASS")
leuk$survivalBinary24Weeks <- ifelse(leuk$time >= 24, 1, 0)

leuk$log_wbc <- log(leuk$wbc)
cell count (log_wbc)
logistic_model <- glm(survivalBinary24Weeks ~ log_wbc + ag, data = leuk, family = binomial)
summary(logistic_model)

par(mfrow = c(2, 2))  # Set up a 2x2 plotting area
plot(logistic_model)

library(ISLR)
data("Weekly")
View(Weekly)
Weekly dataset
summary(Weekly)

# Histogram
hist(Weekly$Today, col = "skyblue", main = "Distribution of Weekly Returns",
xlab = "Today")
# Scatter plot
plot(Weekly$Lag1, Weekly$Today, main = "Lag1 vs Today",
xlab = "Lag1", ylab = "Today", col = "blue", pch = 19)
# Boxplot
boxplot(Weekly$Today ~ Weekly$Direction, main = "Boxplot of Today by Direction",
xlab = "Direction", ylab = "Today", col = c("red", "green"))
#Time series
plot(Weekly$Year, Weekly$Today, type = "l", main = "Time Series of Weekly Returns",
xlab = "Year", ylab = "Today", col = "darkorange")
logistic_model_full <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
+data = Weekly, family = binomial)
summary(logistic_model_full)


predicted_probabilities <- predict(logistic_model_full, type = "response")
predicted_direction <- ifelse(predicted_probabilities > 0.5, "Up", "Down")
confusion_matrix <- table(Predicted = predicted_direction, Actual = Weekly$Direction)
print(confusion_matrix)
```

```
lda_model <- lda(Direction ~ Lag2, data = train_data)
lda_predictions <- predict(lda_model, test_data)
lda_class <- lda_predictions$class
confusion_matrix_lda <- table(Predicted = lda_class, Actual = test_data$Direction)
print("Confusion Matrix (LDA - 2009-2010):")
print(confusion_matrix_lda)
correct_predictions_lda <- sum(diag(confusion_matrix_lda))
total_predictions_lda <- sum(confusion_matrix_lda
accuracy_lda <- correct_predictions_lda / total_predictions_lda
cat("Overall Fraction of Correct Predictions (LDA Accuracy):", accuracy_lda, "\n")
```

```
 qda_model <- qda(Direction ~ Lag2, data = train_data)
qda_predictions <- predict(qda_model, test_data)
qda_class <- qda_predictions$class
confusion_matrix_qda <- table(Predicted = qda_class, Actual = test_data$Direction)
print("Confusion Matrix (QDA - 2009-2010):")
correct_predictions_qda <- sum(diag(confusion_matrix_qda))
total_predictions_qda <- sum(confusion_matrix_qda)
accuracy_qda <- correct_predictions_qda / total_predictions_qda
cat("Overall Fraction of Correct Predictions (QDA Accuracy):", accuracy_qda, "\n")
```