Introduction

The goal of this project is to find three highly correlated stocks, and calculate the daily price change of each by using the given equation $\frac{Y_{t+1}-Y_t}{Y}(Y_t = close\ price\ of\ day\ t)$.

Methodology

1. *Preprocessing*

The following stocks were chosen as the data soruce: *Schlumberger Limited (SLB), Halliburton Company (HAL), and Baker Hughes Incorporated (BHI)*. These companies are the three largest oilfield service companies in the U.S., in the order they are listed above. We believe the prices of these stocks will be correlated because they will be affected by the same factors, specifically, the price and demand for oil. These companies would be similarly affected by new technologies or regulations within the oil industry. Also, it is worth noting that from November 2014 until April 2016 Halliburton and Baker Hughes tentatively planned to merge. Using historical prices from http://finance.yahoo.com, we downloaded the closing prices of these stocks from January 1, 2015 to May 4, 2016, the last 337 trading days. We input the data into one Excel sheet and calculated the daily price change of each using the equation above. This resulted in 336 values for each, which were then separated into a new Excel sheet and converted into a csv file. The file consists of four columns: day, x1 (SLB%change), x2 (HAL%change), x3 (BHI%change).

2. *Data Analysis*

(1) R: We used the *read.csv* command to import the prepared data set, and plotted the data by using the command *plot*(detailed code will be provided in the appendix).

First of all, we used the command *shapiro.test* to determine x1, x2, x3's normality, and the result is that they are all normal. Then we use the command $t.test(, alternative = c("two.sided"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)$ to perform the 1-sample t-tests for x1, x2, and x3. The result are: the mean daily percent change of SLB and BHI are less than 0, while the mean daily percent change for HAL is greater than 0.

Secondly, we performed the ANOVA test between groups by running the code *anova* (detailed steps will be provided in the appendix), and the F-value between group is 0.0891, and the P-value is 0.9148.

Thirdly, in order to tell which two variables are more highly correlated we ran the correlation test between x1 and x2, x1 and x3, x2 and x3 by using the command *cor.test*. Since the correlation between x2 and x3 is 0.854, hence, we believe that x2 and x3 are the most highly correlated, which implies that they are very significant.

After that, we started to determine which one is the dependant variable by creating the linear models when $x1 = x2 + x3(denoted\ as\ model)$, $x2 = x1 + x3(denoted\ as\ model1)$, and $x3 = x2 + x1(denoted\ as\ model2)$. And we applied the code *summary* to find the value of R^2 and the P-value of each model. The P-value and the R^2 value of the model1 is the best of all the models: $R^2 = 0.8176$, $P - value = 2.2e^{-16}$. Therefore, we believe that x2 should be the dependant variable, and x1 and x3 be the independent variables. In order to determine this linear model in the form of $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$,we need to find slopes and the y-intercept, which in our case: y-intercept = estimated intercept = 0.000612, slope of x1 = estimated value of x1 = 0.521737, slope of x3 = estimated value of x3 = 0.545002. Hence, the linear equation will be $y_i = 0.000612 + 0.521737x_{1i} + 0.545002x_{3i}$.

Conclusion

$$y_i = 0.000612 + 0.521737x_{1i} + 0.545002x_{3i}$$

We conclude that for all of these three stocks the mean daily percent change is not significantly different from zero. This is confirmed by our ANOVA test, which shows that our three samples are not significantly different from each other. We found that there was a strong positive correlation between any pair of the three stocks (0.67, 0.79, 0.85). Lastly, we found that using the prices of *SLB* and *BHI* to predict *HAL* was the best model (the equation shown above).The R^2 from this model is 0.8176, showing that it is very predictive. The equation for the model shows that to predict the price of *HAL*, the prices of *SLB* and *BHI* are weighted very similarly.

Appendix

Code used in R:

```r
data <- read.csv(file.choose(), header= TRUE)
View(data)
head(x1)
x1 <- data$x1
x2 <- data$x2
x3 <- data$x3
tapply(x1,x2,summary)
cor.test(x1,x2)
cor.test(x1,x3)
cor.test(x2,x3)
t.test(x1,alternative = c("two.sided"),mu=0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
t.test(x2,alternative = c("two.sided"),mu=0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
t.test(x3,alternative = c("two.sided"),mu=0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
shapiro.test(x1)
shapiro.test(x2)
shapiro.test(x3)
plot <- scatterplot3d(x1,x2,x3,pch=16,highlight.3d = TRUE,type="h",main="3D Scatter Plot with Vertical Lines and Regression Plane")
x = c(x1,x2,x3)
fit = lm(formula = x ~ x1+x2+x3)
anova(fit)
groups = factor(rep(letters[1:3], each = 336))
fit = lm(formula = x ~ groups)
anova(fit)
qchisq(0.950, 2)
model <- lm(x1 ~ x2+x3)
anova (model)
model1 <- lm(x2 ~ x1+x3)
anova (model1)
model2 <- lm(x3 ~ x2+x1)
anova (model2)
```

R plots:

```
> shapiro.test(x1)

        Shapiro-Wilk normality test

data:  x1
W = 0.9855, p-value = 0.001866

> shapiro.test(x2)

        Shapiro-Wilk normality test

data:  x2
W = 0.98699, p-value = 0.004106

> shapiro.test(x3)

        Shapiro-Wilk normality test

data:  x3
W = 0.98859, p-value = 0.009779


        One Sample t-test

data:  x1
t = -0.23197, df = 335, p-value = 0.8167
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.002072493  0.001635249
sample estimates:
    mean of x
-0.0002186221


        One Sample t-test

data:  x2
t = 0.22779, df = 335, p-value = 0.8199
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.002143403  0.002704835
sample estimates:
   mean of x
0.0002807156


data:  x3
t = -0.30008, df = 335, p-value = 0.7643
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.003011471  0.002214267
sample estimates:
    mean of x
-0.0003986016

> groups = factor(rep(letters[1:3], each = 336))
> fit = lm(formula = x ~ groups)
> anova(fit)
Analysis of Variance Table

Response: x
            Df  Sum Sq    Mean Sq F value Pr(>F)
groups       2 0.00008 0.00004162  0.0891 0.9148
Residuals 1005 0.46952 0.00046719
> qchisq(0.950, 2)
[1] 5.991465
```

```
> cor.test(x1,x2)

        Pearson's product-moment correlation

data:  x1 and x2
t = 23.675, df = 334, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7479415 0.8284195
sample estimates:
      cor
0.7915885

> cor.test(x1,x3)

        Pearson's product-moment correlation

data:  x1 and x3
t = 16.419, df = 334, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6045318 0.7235564
sample estimates:
      cor
0.6682995


> cor.test(x2,x3)

        Pearson's product-moment correlation

data:  x2 and x3
t = 30.01, df = 334, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8222362 0.8806133
sample estimates:
      cor
0.8540921

> summary(model)

Call:
lm(formula = x1 ~ x2 + x3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.050384 -0.006054  0.000015  0.005871  0.042905

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0004020  0.0005783  -0.695    0.487
x2           0.6241845  0.0492207  12.681   <2e-16 ***
x3          -0.0204311  0.0456651  -0.447    0.655
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01058 on 333 degrees of freedom
Multiple R-squared:  0.6268,    Adjusted R-squared:  0.6246
F-statistic: 279.7 on 2 and 333 DF,  p-value: < 2.2e-16
```

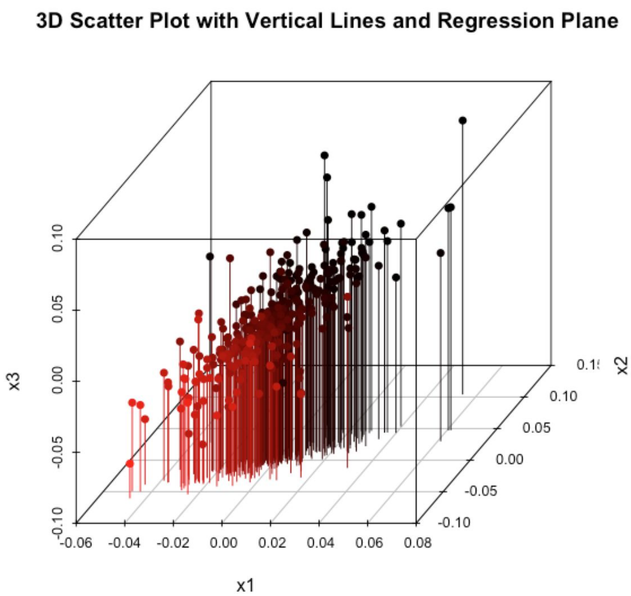
```
> summary(model1)

Call:
lm(formula = x2 ~ x1 + x3)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.039945 -0.005226 -0.000267  0.005433  0.038739 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.000612   0.000528   1.159    0.247    
x1          0.521737   0.041142  12.681   <2e-16 ***
x3          0.545002   0.029191  18.670   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009677 on 333 degrees of freedom
Multiple R-squared:  0.8176,	Adjusted R-squared:  0.8165 
F-statistic: 746.2 on 2 and 333 DF,  p-value: < 2.2e-16
```

```
> summary(model2)

Call:
lm(formula = x3 ~ x2 + x1)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.061244 -0.005413  0.000431  0.006042  0.045917 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0006685  0.0006933  -0.964    0.336    
x2           0.9383956  0.0502616  18.670   <2e-16 ***
x1          -0.0294048  0.0657220  -0.447    0.655    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0127 on 333 degrees of freedom
Multiple R-squared:  0.7296,	Adjusted R-squared:  0.728 
F-statistic: 449.3 on 2 and 333 DF,  p-value: < 2.2e-16
```

Outlier test for the selected model:

```
> outlier.test(model1)
     rstudent unadjusted p-value Bonferonni p
267 -4.263444         2.6270e-05    0.0088266
315  4.155360         4.1364e-05    0.0138980
```

3-D Scatter plot of x1, x2, x3:



3D Scatter Plot with Vertical Lines and Regression Plane
x3
0.10
0.05
0.00
-0.05
-0.10
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06
0.08
x1
x2
0.10
0.05
0.00
-0.05
-0.10