Research Project

Load Packages

```
library(quantmod)
library(tidyverse)
library(XLConnect)
library(tidyquant)
library(tseries)
library(reshape2)
library(readxl)
```

Function: Calculate monthly return

```
get_return <- function(symbol, start, end, compression = 'm'){
        price <- get.hist.quote(instrument = symbol, start = start, end = end, compression = compressio
        Date <- index(price)

        price <- (price %>% as.data.frame()
                  %>% mutate(Date = Date)
                  %>% select(Date, everything())
        )

        returns <- price %>% mutate(returns = Close/lag(Close) - 1) %>% select(-Close) %>% filter(compl

        returns <- returns %>% as.tibble()
        return(returns)
}
```

Data range and query stock

```
start <- '2012-06-01'

end <- '2018-06-02'

query_stock <- 'APA'
```

I. Carhart four-factor model

1) Query stock returns

```
returns <- get_return(query_stock, start, end)

returns$Date <- as.yearmon(returns$Date)
```

2) Data: Fama French 3 factors

```
ff_factors <- read.csv("F-F_Research_Data_Factors.CSV", header = TRUE, skip = 3, nrows = 1102)

colnames(ff_factors)[1] <- 'Date'
```

```r
# Format Date into yearmon

ff_factors$Date <- as.yearmon(as.Date(paste(ff_factors$Date, '01', sep = ''), format='%Y%m%d'))

ff_factors[,-1] <- ff_factors[,-1]/100
```

3) Data: Fama French Momentum

```r
ff_mom <- read.csv("F-F_Momentum_Factor.CSV", header = TRUE, skip = 13, nrows = 1095)

colnames(ff_mom)[1] <- 'Date'

# Format Date into yearmon

ff_mom$Date <- as.yearmon(as.Date(paste(ff_mom$Date, '01', sep = ''), format='%Y%m%d'))

ff_mom[,-1] <- ff_mom[,-1]/100
```

4) Run Regression

```r
# Merge stock returns, ff_mom, ff_factors by Date

df <- Reduce(function(x,y) merge(x, y, by = 'Date'), list(ff_factors, ff_mom, returns))

# Calculate excess stock return

df <- df %>% mutate(Ex.return = returns - RF)

# Run regression

df <- df %>% select(-c(RF, returns))

summary(lm(Ex.return ~ ., data = df[,-1]))
```

```
##
## Call:
## lm(formula = Ex.return ~ ., data = df[, -1])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.199535 -0.036759 -0.003474  0.047167  0.281671
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01179    0.01110  -1.062 0.292319
## Mkt.RF       0.71625    0.37029   1.934 0.057504 .
## SMB          0.52665    0.45071   1.169 0.246939
## HML          0.37236    0.52553   0.709 0.481185
## Mom         -1.49069    0.40479  -3.683 0.000476 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.083 on 64 degrees of freedom
```

```
## Multiple R-squared:  0.3652, Adjusted R-squared:  0.3255
## F-statistic: 9.206 on 4 and 64 DF,  p-value: 6.127e-06
```

>> According to the Carhart four-factor model, momentum and market factor are significant.

II. PCA Analysis

1) Filter out stocks in the same sub-sector with query_stock. (Data: GICS Subsector)

```
sub_industry <- read_xlsx("SP500 Sectors.xlsx", sheet = 1)
```

2) Sub-industry for query stock

```
query_industry <- sub_industry %>% filter(`Ticker symbol` == query_stock) %>% select(`GICS Sub Industry

query_industry %>% as.character()
```

```
## [1] "Oil & Gas Exploration & Production"
```

3) Filter out peers

```
peers <- sub_industry %>% filter(`GICS Sub Industry` == query_industry) %>% select(`Ticker symbol`)

colnames(peers) <- 'symbol'

peers %>% unlist() %>% as.vector()
```

```
##  [1] "APA" "APC" "COG" "COP" "CXO" "DVN" "EOG" "EQT" "MRO" "NBL" "NFX"
## [12] "OXY" "PXD" "XEC"
```

4) Calculate returns for stocks in the selected sub-sector

```
peers_return <- peers %>% mutate(returns = map(symbol, function(.x) get_return(.x, start, end)))

peers_return <- peers_return %>% unnest()

peers_return <- dcast(peers_return, Date ~ symbol)

# Exclude stocks with shorter history than the data range selected
peers_return <- peers_return[, colSums(is.na(peers_return)) == 0]
```

5) PCA analysis

```
pca <- prcomp(peers_return[,-1])
```

6) Select retained PCs based on Scaled Average Eigenvalues (Keep eigenvalues larger than 0.7*mean(eigenvalues))

```
keep_scale <- mean(pca$sdev^2)*0.7
n <- sum(pca$sdev^2 > keep_scale)
n
```

```
## [1] 3
```

7) Regress APA stock return on PCs

```r
pcs <- pca$x[,c(1:n)]

pc_tbl <- peers_return %>% select(Date) %>% cbind(pcs)

pc_tbl$Date <- as.yearmon(pc_tbl$Date)

tbl <- merge(returns, pc_tbl)

summary(lm(returns ~., data = tbl[,-1]))
```

```
##
## Call:
## lm(formula = returns ~ ., data = tbl[, -1])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.103837 -0.040021  0.000003  0.030655  0.208151
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.004374   0.006746  -0.648   0.5189
## PC1          0.306015   0.025415  12.041   <2e-16 ***
## PC2          0.011107   0.071625   0.155   0.8772
## PC3          0.185544   0.081277   2.283   0.0256 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05724 on 68 degrees of freedom
## Multiple R-squared:  0.6884, Adjusted R-squared:  0.6746
## F-statistic: 50.07 on 3 and 68 DF,  p-value: < 2.2e-16
```

>> PC1 and PC3 are significant, R-squared is 0.6884.

8) Variance explained by PCA

```r
summary(pca)$importance[,1:3]
```

```
##                              PC1        PC2        PC3
## Standard deviation     0.2673055 0.09485103 0.08358663
## Proportion of Variance 0.6443500 0.08113000 0.06301000
## Cumulative Proportion  0.6443500 0.72548000 0.78849000
```

>> First three PCs have explained 78.85% of variation

9) PC1, PC2, PC3 loadings

```r
pca$rotation[,1:n]
```

```
##           PC1         PC2         PC3
## APA 0.30601460  0.01110738  0.18554418
## APC 0.28173368  0.02532967  0.18224084
## COG 0.08428241  0.69213010 -0.25283826
## COP 0.22934632  0.01654187  0.25576330
## CXO 0.24475107 -0.22934286 -0.36795702
## DVN 0.38876059  0.16541424  0.18130716
## EOG 0.24303301 -0.06536542 -0.13354567
## EQT 0.13371813  0.43954872 -0.40522715
## MRO 0.41485058  0.23304398  0.43195885
## NBL 0.24599763 -0.11200576 -0.09309682
## NFX 0.31873719 -0.31202004 -0.05127994
## OXY 0.15122415 -0.13066884 -0.09247165
## PXD 0.24967147 -0.24122999 -0.23158432
## XEC 0.24470525 -0.05480925 -0.43891234
```

>> All stocks have positive weights in PC1, check its correlation with Oil Price

10) Data: WTI

```
WTI <- read.csv('WTI.csv', header = TRUE)

WTI$Date <- as.yearmon(as.Date(WTI$Date, format='%m/%d/%Y'))

WTI <- WTI %>% mutate(WTI.return = WTI/lag(WTI)-1) %>% select(-WTI)
```

11) Correlation tabel

```
df_new <- merge(tbl, WTI)

cor(df_new[,-1])
```

```
##                returns          PC1          PC2          PC3 WTI.return
## returns    1.000000000  0.820202314  0.001973072  0.146820534  0.3313157
## PC1        0.820202314  1.000000000 -0.001632007 -0.001838155  0.4475515
## PC2        0.001973072 -0.001632007  1.000000000 -0.003827019  0.1558971
## PC3        0.146820534 -0.001838155 -0.003827019  1.000000000  0.0875003
## WTI.return 0.331315750  0.447551489  0.155897139  0.087500301  1.0000000
```

>> The correlation betwwen PC1 and WTI is 0.44, because the sub-industry of APA is Oil & Gas Exploration & Production, it may have higher correlation with Oil & Gas index, commodity index or PMI. (But those data cannot be downloaded from website, it is not feasible to verify this guess)

III. Further thoughts

>> Because this poject is to study the return of one single stock, a model based on valuation variables would explain more firm-specific returns, factors including Size, P/B, ROE, Dividends per share, Debt/Price, etc. But those historical valuation data cannot be obtained online, so it is not feasible to do more research on this idea.