Trading ETF’s, Daniel Emil Wiinberg S133232

```
setwd("/Users/dwiinberg/Desktop/Project 1")
```

```
D = read.table("finans1_data.csv", header=TRUE, sep=";", as.is=TRUE)
D = D[ ,c("t","AGG","VAW","IWN","SPY")]
```

Descriptive analysis

The raw data includes weekly returns for 95 ETFs, with the first week being 2006-5-5 and the last week being 2015-5-8 (date format is yyyy-mm-dd). In total this is 454 data points for every ETF.

From this raw data four ETFs are selected, they are:

- AGG - iShares Core Total US Bond Market ETF
- VAW - Vanguard Materials ETF
- IWN - iShares Russell 2000 Value Index Fund
- SPY - SPDR Trust, Series 1 is a unit investment trust

```
head(D)
```

```
##           t          AGG         VAW          IWN          SPY
## 1  2006-5-5 -0.006088280  0.03157274  0.023603147  0.007986613
## 2 2006-5-12 -0.003675345 -0.02745995 -0.048071912 -0.024750981
## 3 2006-5-19  0.006660518 -0.05000000 -0.014096072 -0.016558341
## 4 2006-5-26  0.001832248  0.01253870  0.009716824  0.010070810
## 5  2006-6-5 -0.004775452 -0.02232075 -0.017734396 -0.009814613
## 6 2006-6-12  0.001225115 -0.06098514 -0.038348495 -0.024622404
```

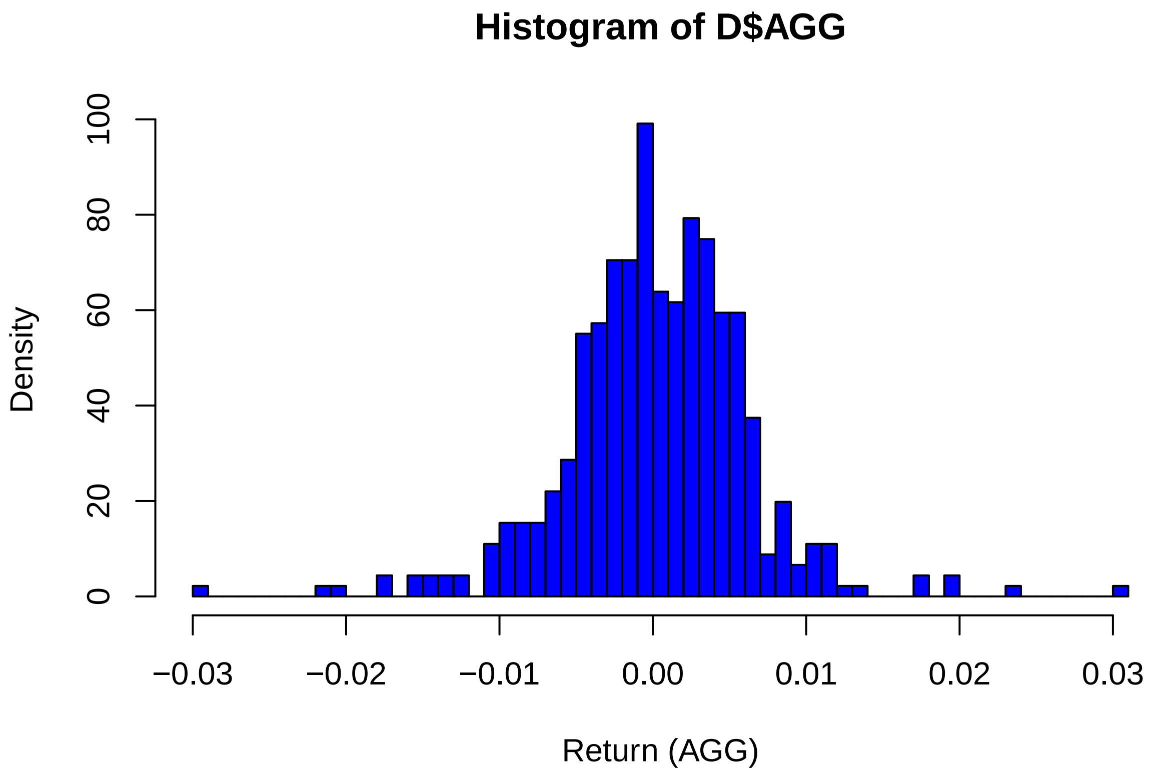
From the rows we can se that we are dealing with quantitative data (the return of the ETF) and the week this return was given.

Looking through the data is can be seen that the data is complete for the four ETF’s selected. However tt can also be seen that SPY has an outlier of -18, which looks to be wrong in comparison to the rest of the data. But since it’s just one datapoint out of 454, it won’t have much of an impact.

Histogram

Shown is a histogram of the empirical density of weekly return from the AGG fund.

```
hist(D$AGG, xlab="Return (AGG)", prob=TRUE, breaks = 50, col="blue")
```



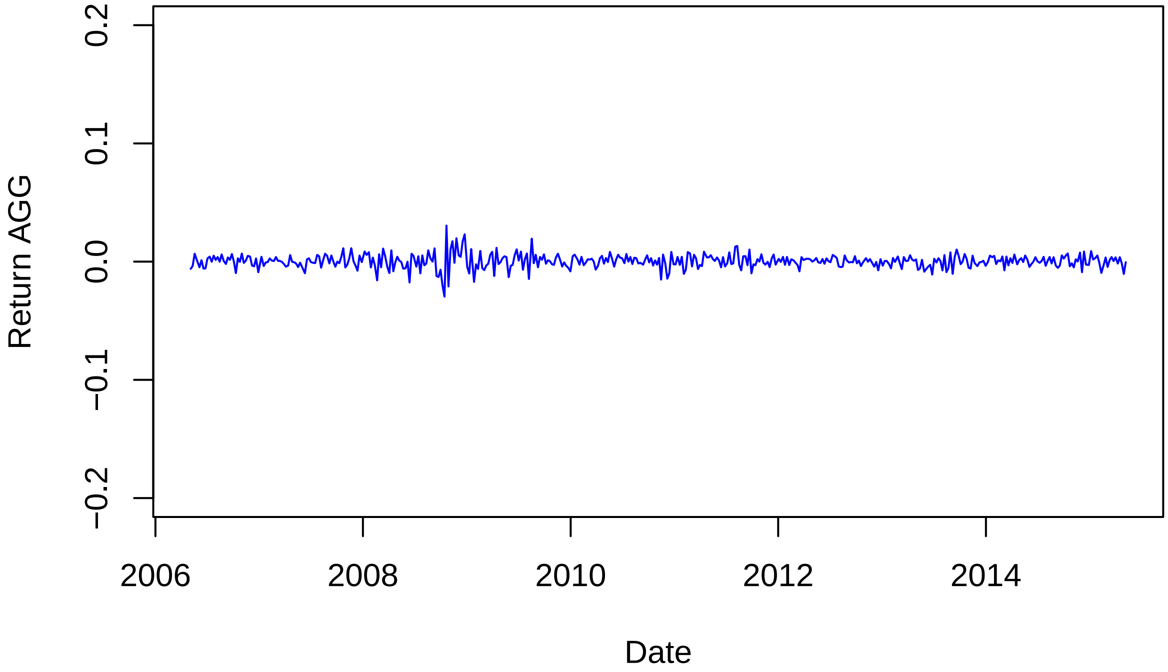
From the histogram it can be seen that the data is largely symmetrical, with both positive and negative returns, as expected from the stock market. The variations are between -0.03 and 0.03, with the largest quantity being between -0.01 and 0.01.

Return over time

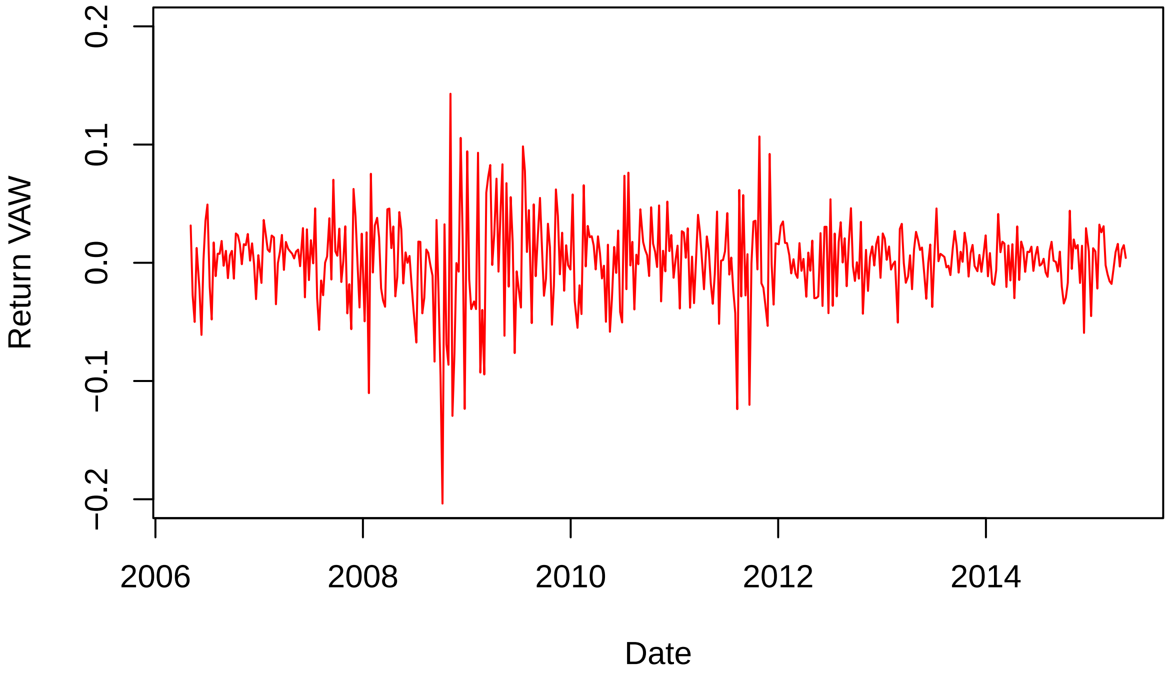
```
#Setting the t-variable as a date variable.
D$t = as.Date(x=D$t, format="%Y-%m-%d")
```

Next we plot the weekly return over time for each of the four ETF's.

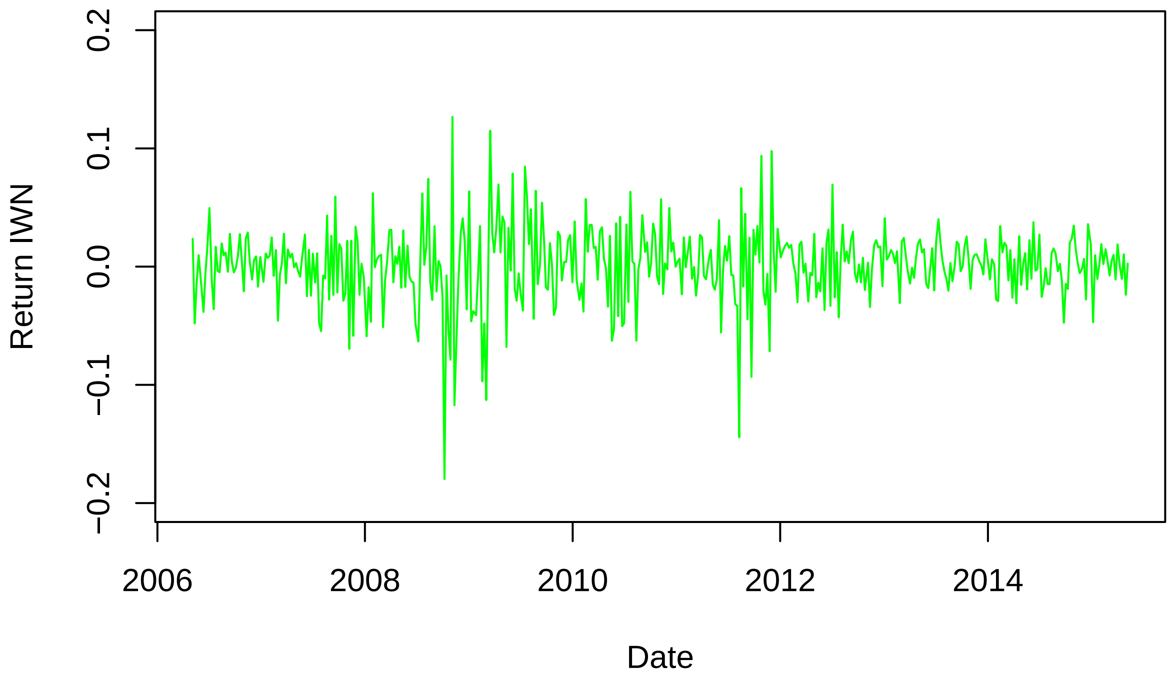
```
ylim = c(-0.2,0.2)
plot(D$t, D$AGG, type="l", ylim=ylim, xlab="Date", ylab="Return AGG", col="blue")
```



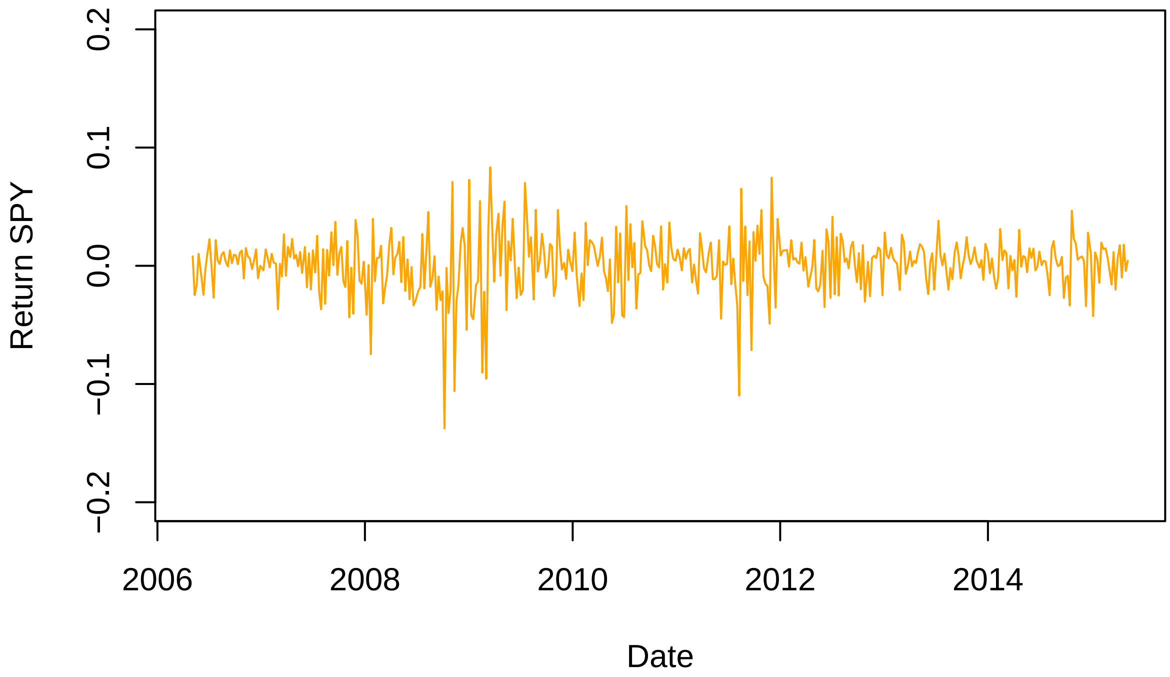
```
plot(D$t, D$VAW, type="l", ylim=ylim, xlab="Date", ylab="Return VAW", col="red")
```



```
plot(D$t, D$IWN, type="l", ylim=ylim, xlab="Date", ylab="Return IWN", col="green")
```



```
plot(D$t, D$SPY, type="l", ylim=ylim, xlab="Date", ylab="Return SPY", col="orange")
```



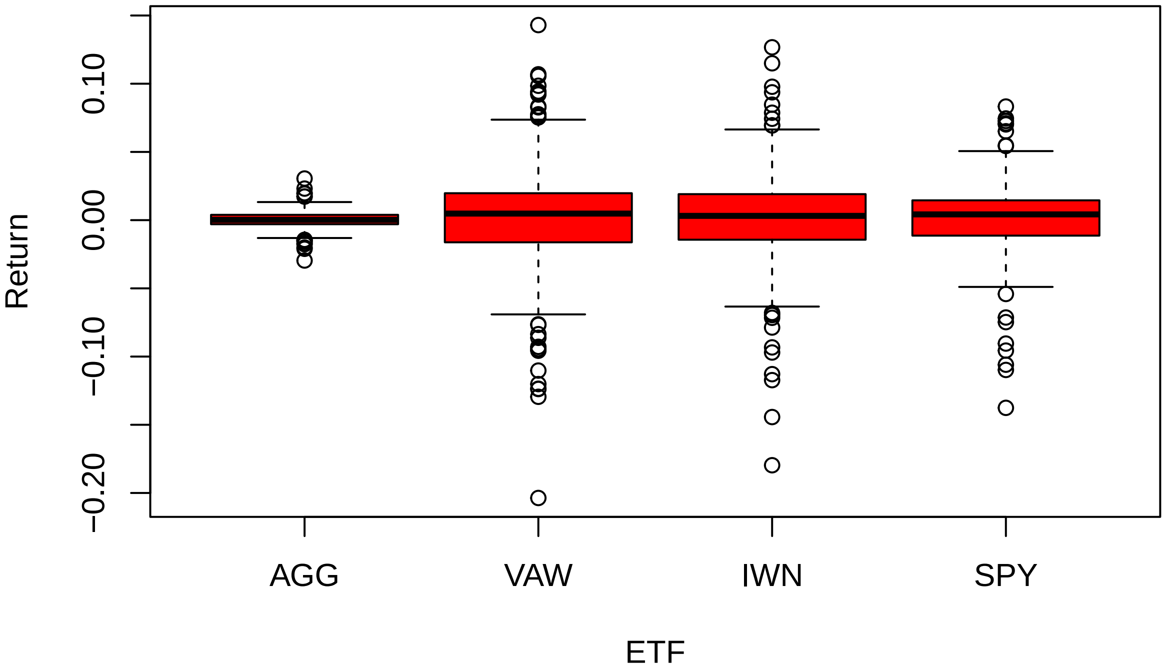
From the plots it can be seen that they are clearly very different in the return they provide. The least fluctuation can be seen in the AGG fund, which is an ETF tracking the US bond market. Therefore it makes sense that is has less fluctuation.

Besides that it can be seen for the other 3 funds that they have the most fluctuation around 2009 and again in the last quarter of 2011. Otherwise the weekly return fluctuates between -0.1 and 0.1 or less.

Box plots

Creating box plots for the four ETF's.

```
boxplot(D$AGG, D$VAW, D$IWN, D$SPY, names=c("AGG", "VAW", "IWN", "SPY"),
        xlab="ETF", ylab="Return", col="red")
```



The box plots are all largely symmetrical, with a mean roughly around zero (which will be explored more in detail next). The horizontal lines mark outliers for each ot the plots, from them it can be that values larger or smaller than 0.01 is not common. It can also be seen that VAW has the largest deviation from the mean, followed by IWN, then SPY and AGG has the least fluctuation.

```
#Calculation number of observations, sample mean and variance for the four ETF's.

# Code below has been commented out to avoid clutter in the document. Results
# has been been filled out in the table.

# sum(!is.na(D$AGG))
# var(D$AGG, na.rm=TRUE)
# sd(D$AGG, na.rm=TRUE)
# summary(D$AGG, na.rm=TRUE)
#
# sum(!is.na(D$VAW))
# var(D$VAW, na.rm=TRUE)
# sd(D$VAW, na.rm=TRUE)
# summary(D$VAW, na.rm=TRUE)
#
# sum(!is.na(D$IWN))
# var(D$IWN, na.rm=TRUE)
# sd(D$IWN, na.rm=TRUE)
# summary(D$IWN, na.rm=TRUE)
#
# sum(!is.na(D$SPY))
# var(D$SPY, na.rm=TRUE)
# sd(D$SPY, na.rm=TRUE)
# summary(D$SPY, na.rm=TRUE)
```

ETF	Number of obs.	Sample mean	Sample variance	Std. dev.	Lower quantile	Median	Upper quantile
	n	(x)	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
AGG	454	0.0002658	3.571e-05	0.005976	-0.002973	0.0002374	0.003893
VAW	454	0.001794	0.001302	0.03608	-0.01610	0.004798	0.019685
IWN	454	0.001188	0.001025	0.03202	-0.01431	0.003120	0.019056
SPY	454	0.001360	0.0006143	0.02479	-0.01133	0.004216	0.014498

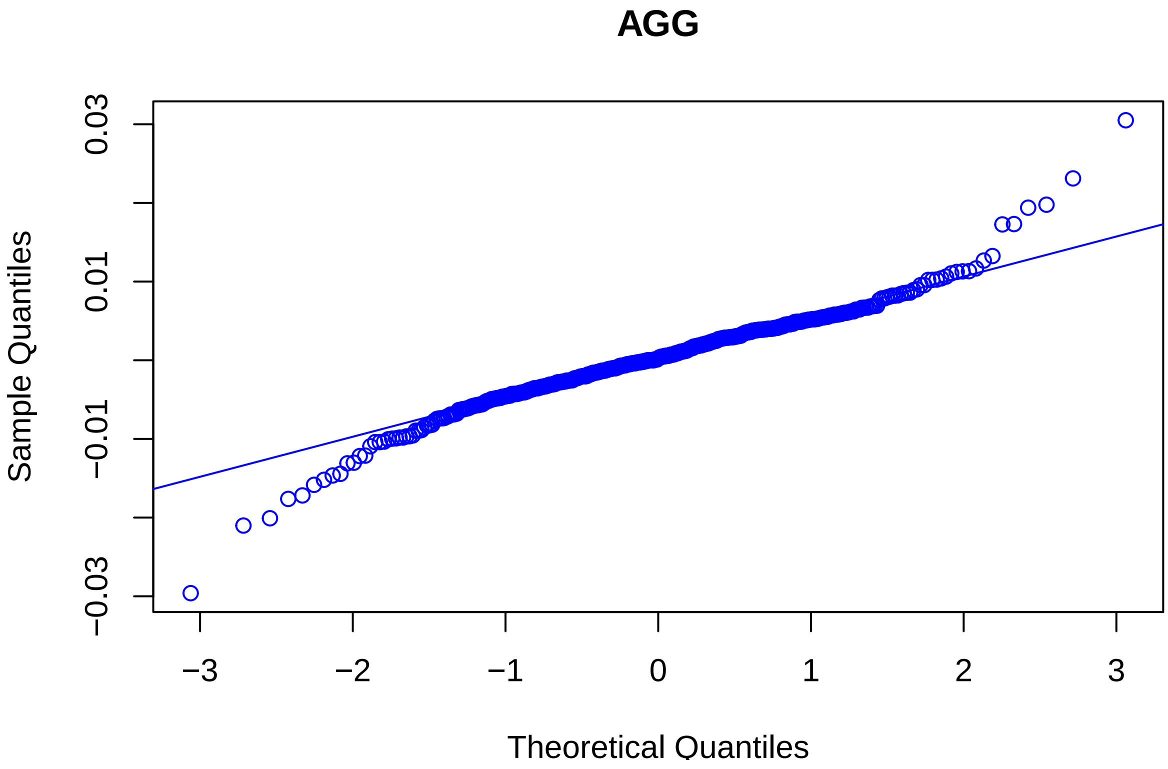
Table 1: Summary statistics

From the table the quantiles can be read more precisely than from the box plot, also the mean, variance and standard deviation can be seen.

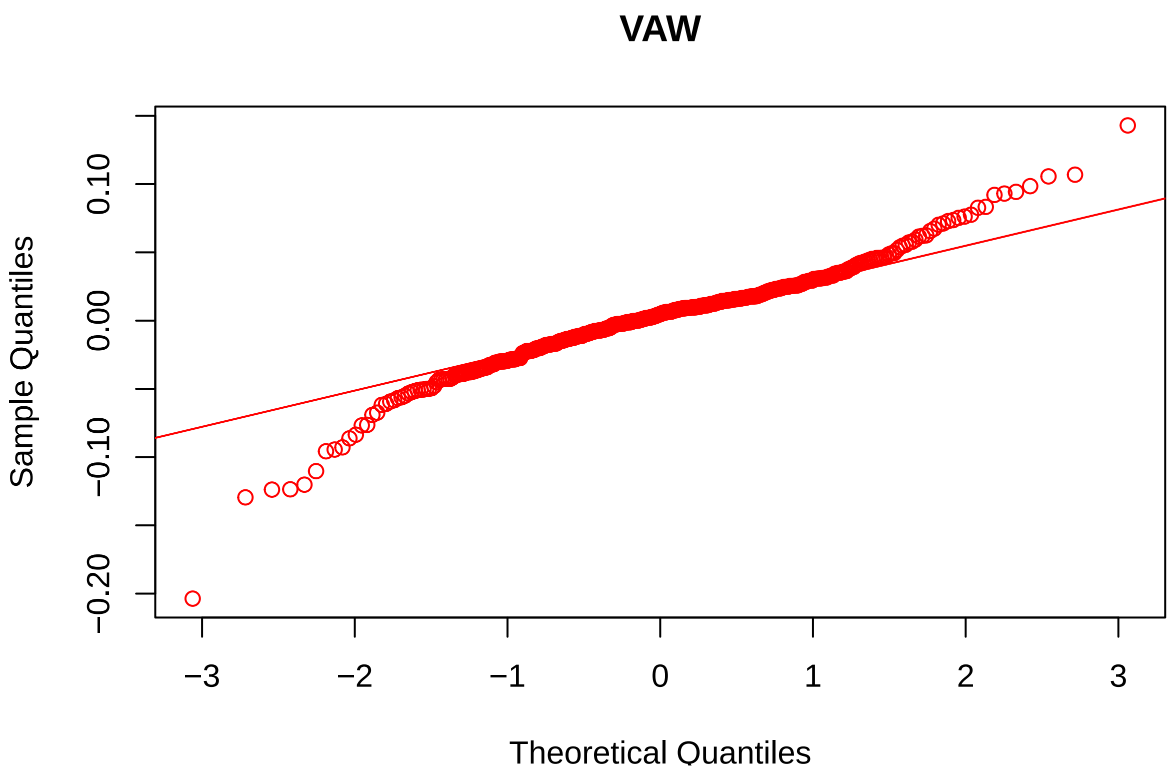
Statistical analysis

Below are QQ plots for the four ETF's.

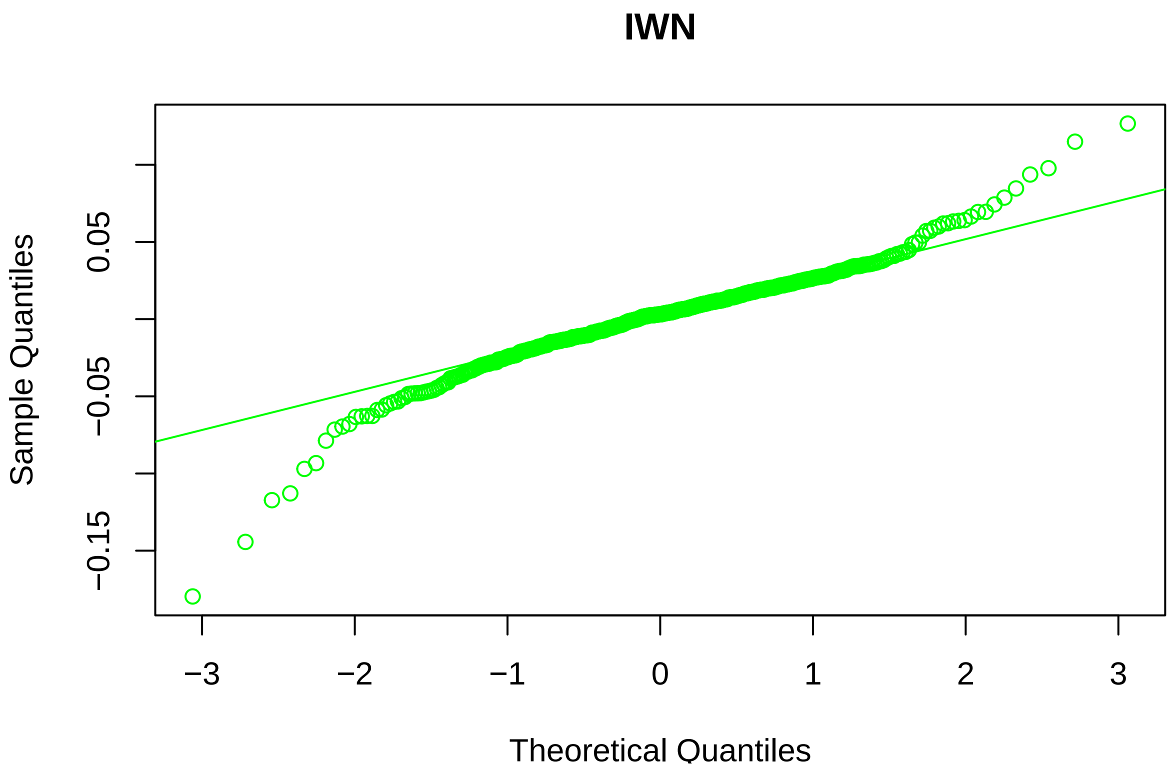
```
qqnorm(D$AGG, main="AGG", col="blue")
qqline(D$AGG, col="blue")
```



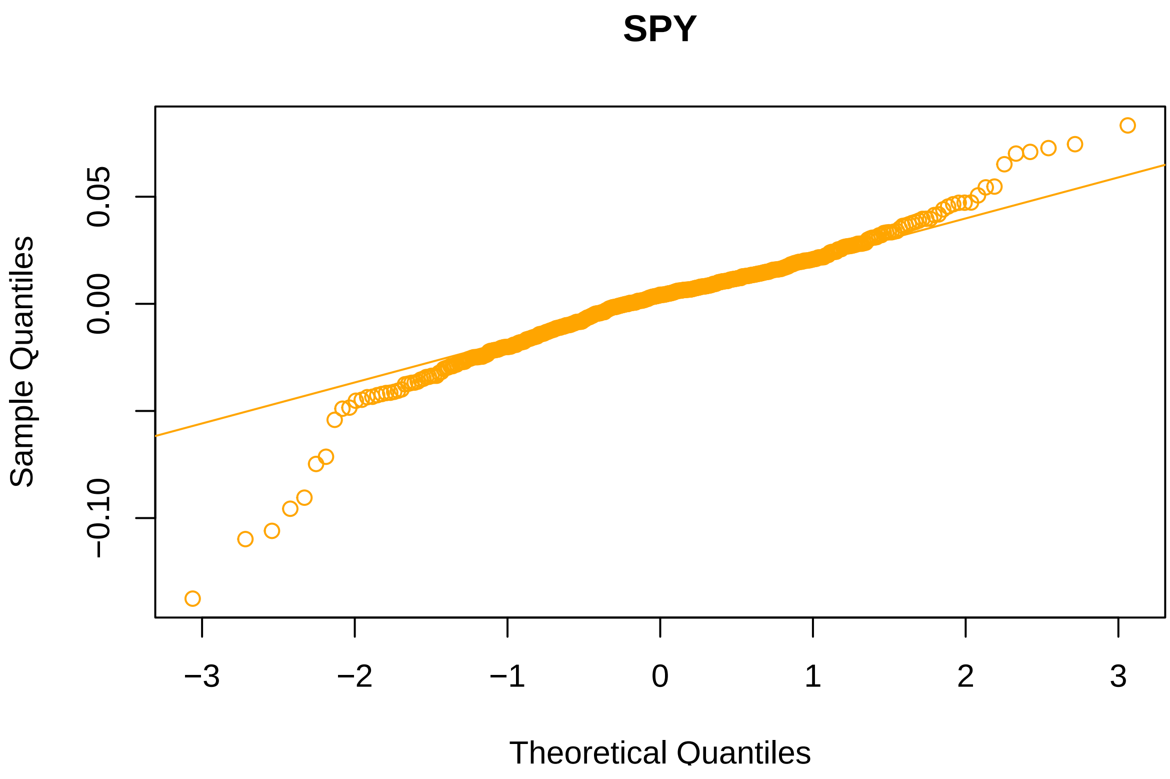
```
qqnorm(D$VAW, main="VAW", col="red")
qqline(D$VAW, col="red")
```



```
qqnorm(D$IWN, main="IWN", col="green")
qqline(D$IWN, col="green")
```



```
qqnorm(D$SPY, main="SPY", col="orange")
qqline(D$SPY, col="orange")
```



From the plots it can be clearly seen that they mostly follow a normal distribution. However the data also show that the tails of the data has a greater variation than what would be expected of a true normal distribution. In the investment world this is known as "fat tails", which are returns that are wildly higher or lower than expected.

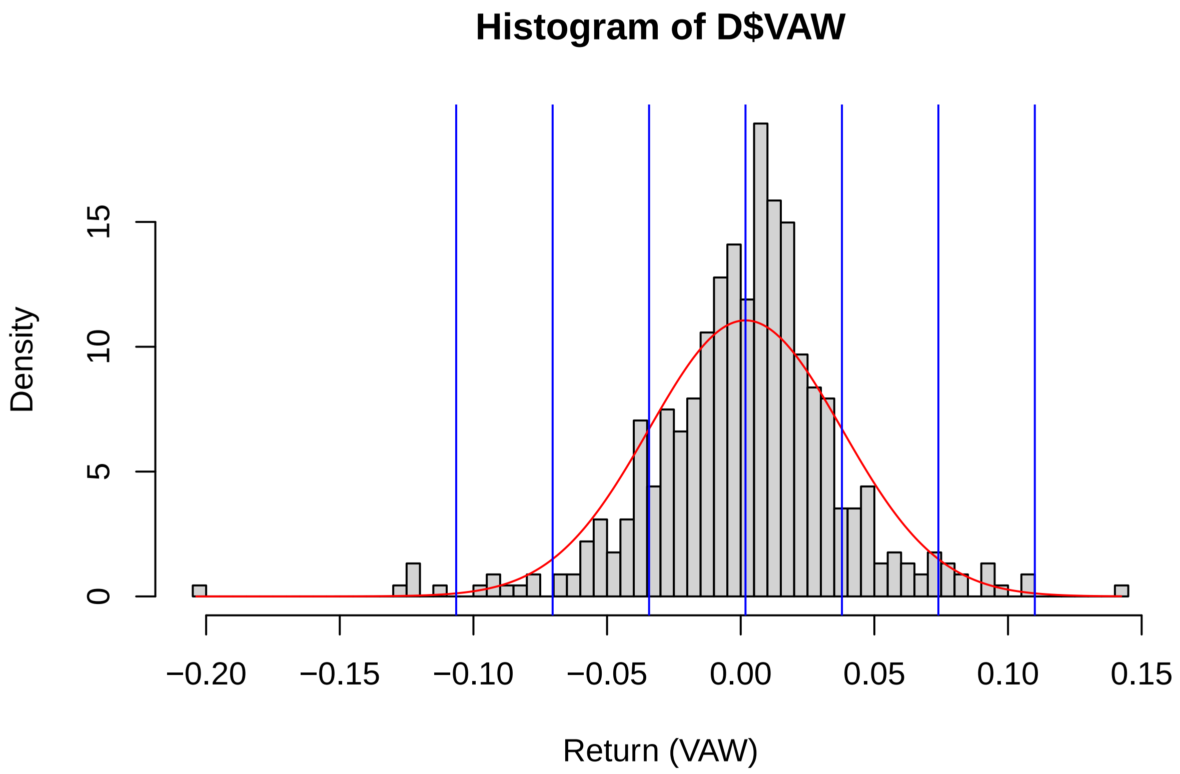
Let's examine the VAW ETF, which has the highest mean return and standard deviation of the funds.

```
#VAW: expected weekly return = 0.1794%, standard deviation = 3.608%
mean_vaw = 0.001794
sd_vaw = 0.03608

hist(D$VAW, xlab="Return (VAW)", prob=TRUE, breaks = 100)

#Theoretical normal distribution
x = seq(from = min(D$VAW), to = max(D$VAW), by=0.001)
y = dnorm(x, mean_vaw, sd_vaw)
lines(x, y, col="red")

#Vertical lines highlighting 1,2,3 standard deviations to each side and the mean.
for (i in seq(-3, 3, 1)) {
  abline(v = mean_vaw + i*sd_vaw, col="blue")
}
```



From this plot it can be quite clearly seen that there are a few very high returns but also a few very big losses, which both are quite surprising for a normal distribution. Let's examine the probability of the biggest loss:

```
loss_vaw = min(D$VAW)
loss_vaw
```

```
## [1] -0.2036603
```

```
#Number of standard distributions out:
distributions_out = loss_vaw/sd_vaw
distributions_out
```

```
## [1] -5.644687
```

```
#So the biggest loss of 20.4% in a week is almost 6 standard deviations out!

#The probability if this happening in a standard deviation is:
pnorm(loss_vaw, mean = mean_vaw, sd = sd_vaw)
```

```
## [1] 6.189954e-09
```

Obviously the chance of this event occurring is ridiculously small, but even though this is not an uncommon occurrence in investing. Even though, when reading about the subject it is widely considered that the normal distribution is the best available model. Maybe a lack of data or other factors plays into the results, or maybe there is some other better model that has not been found yet.

Confidence interval

The formula for the one-sample mean confidence interval is:

$$\overline{x} = t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

	Lower bound of CI	Upper bound of CI
AGG	-0.0002854	0.0008169
VAW	-0.001534	0.005122
IWN	-0.001765	0.004141
SPY	-0.0009260	0.003646

Table 2: Confidence intervals

```
#Calculated "by hand":
mean_agg = mean(D$AGG)
sd_agg = sd(D$AGG)
mean_agg + c(-1, 1) * qt(0.975, length(D$AGG) - 1) * (sd_agg / sqrt(length(D$AGG)))
```

```
## [1] -0.0002854073  0.0008169213
```

```
#Using built-in R function
t.test(D$AGG, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  D$AGG
## t = 0.94757, df = 453, p-value = 0.3439
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.0002854073  0.0008169213
## sample estimates:
##   mean of x
## 0.000265757
```

```
#t.test(D$VAW, conf.level=0.95)
#t.test(D$IWN, conf.level=0.95)
#t.test(D$SPY, conf.level=0.95)
```

As it can be seen, the values calculated “by hand” are the same as from the t.test functionality in R.

Below can be seen the confidence intervals calculated for the other ETF’s.

Null-hypothesis test

In general we say that the null-hypothesis is rejected if the p-value is smaller than 0.05. The smaller the p-value, the stronger the evidence against H_0.

```
#Using formula 3.23 to calculate the P-value
t_obs = (mean_agg - 0) / (sd_agg / sqrt(length(D$AGG)))
2 * (1 -pt(t_obs, df=length(D$AGG)-1))
```

```
## [1] 0.3438511
```

```
#Same results as before, but this time we're specifying mu. Doesn't make a difference though.
t.test(D$AGG, mu=0)
```

```
##
##  One Sample t-test
##
## data:  D$AGG
## t = 0.94757, df = 453, p-value = 0.3439
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.0002854073  0.0008169213
## sample estimates:
##   mean of x
## 0.000265757
```

From the results it can be seen that the manual calculation is correct when compared to the built-in R function. It can also be seen that with a P-value = 0.344 there is little or no evidence against H_0 (consulting table 3.1 in the course notes).

This can also be seen from the confidence intervals, since they describe the interval in which the true mean might be. Since the lower bound for the confidence interval for AGG (and the other ETF's for that matter) is below zero, it cannot be determined if the actual mean might actually be zero.

Welch two-sample test

```r
#method 3.51 p 172, using theorem 3.50
n_agg = length(D$AGG); n_vaw = length(D$VAW)
#mean_agg; sd_agg; n_agg
#mean_vaw; sd_vaw; n_vaw

t_obs = ( (mean_vaw - mean_agg) - 0) / sqrt( (sd_agg^2)/n_agg + (sd_vaw^2)/n_vaw )
v = ( (sd_agg^2/n_agg) + (sd_vaw^2/n_vaw) )^2 /
  ( (sd_agg^2/n_agg)^2 / (n_agg-1) + (sd_vaw^2/n_vaw)^2 / (n_vaw-1) )

t_obs; v; 2*(1 - pt(t_obs, v))
```

```
## [1] 0.8903836
```

```
## [1] 477.8351
```

```
## [1] 0.373708
```

```r
t.test(D$VAW, D$AGG)
```

```
##
##  Welch Two Sample t-test
##
## data:  D$VAW and D$AGG
## t = 0.89019, df = 477.83, p-value = 0.3738
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001844827  0.004900893
## sample estimates:
##   mean of x   mean of y
## 0.001793790 0.000265757
```

It can be seen that the values between manual calculation and built-in R functions are the same. The test was necessary to perform since the confidence intervals were overlapping, and the result of the test might have shown that the null-hypothesis could have been rejected anyway. However, since the P-value is (significantly) larger than 0.05, it can't.

Correlation

The formula for the sample correlation coefficient (r) and the sample covariance (s_{xy}):

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

```
#Values for the covariance formula:
cov(D$VAW, D$IWN); sd(D$VAW); sd(D$IWN)
```

```
## [1] 0.0009838237

## [1] 0.03608286

## [1] 0.03201547
```

```
r = (0.0009838237) / (0.03608286 * 0.03201547)
r
```

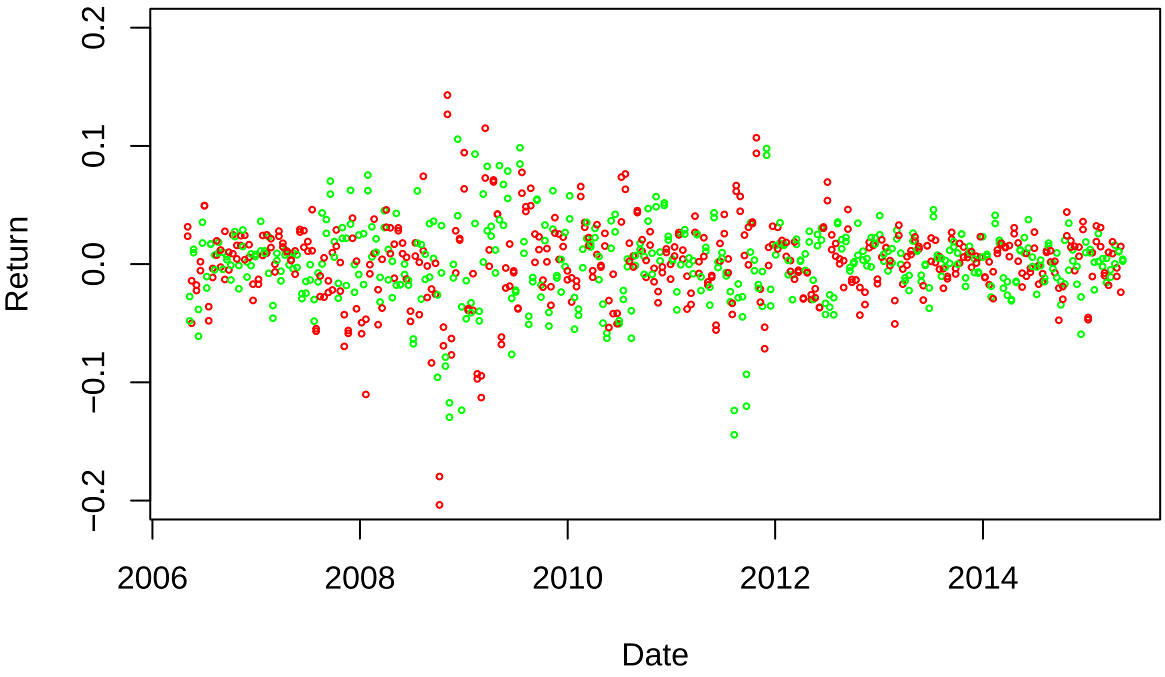
```
## [1] 0.8516408
```

So the covariance between the two ETF's is 0.8516408. When comparing to the R function, it can be seen that it is the same calculated value.

```
# Computing the correlation between selected ETF's
cor(D[ ,c("AGG","VAW","IWN","SPY")], use="pairwise.complete.obs")
```

```
##            AGG        VAW        IWN        SPY
## AGG  1.0000000 -0.1975679 -0.1352621 -0.2187164
## VAW -0.1975679  1.0000000  0.8516407  0.8863608
## IWN -0.1352621  0.8516407  1.0000000  0.9100966
## SPY -0.2187164  0.8863608  0.9100966  1.0000000
```

```
ylim = c(-0.2,0.2)
plot(c(D$t, D$t), c(D$VAW, D$IWN), type="p", ylim=ylim, xlab="Date",
     ylab="Return", col=c("red", "green"), cex=0.4)
```



From the scatterplot it can be seen that if we drew a trendline between the two data sets, the points would fluctuate relatively evenly around that straight line. This is also what we would expect from a covariance of 0.85.