Project: Trading with ETF

Formalities, structure and expectations for the first mandatory project

The assignment consists of two parts. The first part focuses on descriptive analysis of the data. The second part is primarily about confidence intervals and hypothesis tests.

The assignment is formulated in such a way that it can be solved in small "easy" steps. In practice, the assignment must be solved using the statistical software R. Some R code is provided in order to make it easy to get started with the project. However, the code is not complete, and you are encouraged to explore new features in R while working on the project. For example, you could add suitable titles to the plots, or use R's built-in functions for computing confidence intervals and testing hypotheses.

The results of the analysis must be documented in the report using tables, figures, mathematical notation, and explanatory text. Relevant figures and tables must be included within the text, not in the appendix. Present the results of your analysis as you would when explaining them to one of your peers.

Divide the report into subsections, one for each of the questions to be answered.

The report must be handed in as a pdf file. R code should not be included in the report itself but must be handed in as an appendix (a .R-file). The report and appendix must be handed in under Opgaver/Assignments' on CampusNet.

The report text should not exceed 9 pages (excluding figures, tables, and the appendix). A normal page contains 2400 characters.

Figures and tables cannot stand alone - it is important that you describe and explain the R output in words.

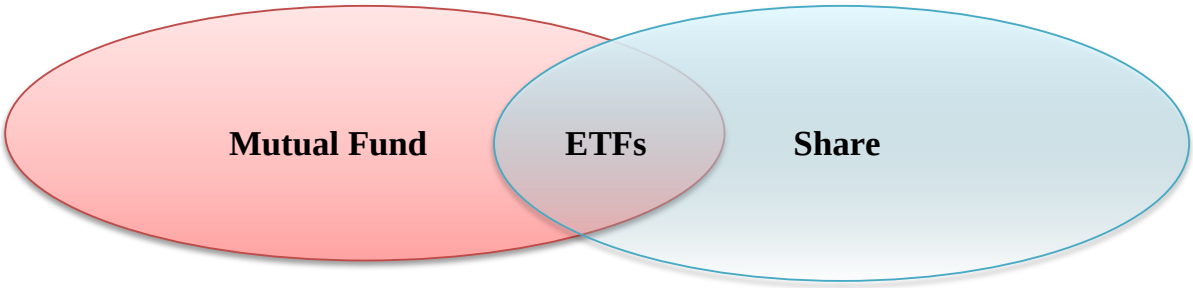
Figures and tables are not included in the assessment of the length of the report. However, it is not in itself an advantage to include many figures, if they are not relevant!

You may work together in groups, but the report must be written individually.

The problem

In this project, the weekly returns for a selection of ETFs are analyzed and modelled. An ETF (Exchange Traded Fund) can be described as a structured publicly traded pool of shares. ETFs are bought and sold in the same way as ordinary shares on a stock exchange.

An ETF is a pooled investment fund similar to a unit trust or mutual fund. For investors, ETFs blend the benefits of pooled funds and shares.



Figur .1: Illustration of the ETF concept.

If you buy for example a simple ETF covering the SP100 Index in the United States it is equivalent to owning a part of all 100 stocks in the index. Thus you avoid buying 100 individual securities and can instead just buy a single.

There are many different ETFs - actually the ETF market is under explosive development. Various strategies are available - for example passive and active approach under which EFTs are administered.

An ETF with passive strategy seek to track the underlying index return as close as possible. Such an ETF is called an index fund. An example would be the EURO STOXX Index of leading Eurozone company shares. This means that the aim of an ETF is to provide investors with the same return as the underlying market. For example, if the EURO STOXX50 Index goes up by 10% during a year, an ETF tracking this index aims to provide investors with the same return, minus fees, which in the case of an ETF is called the Total Expense Ratio (TER). To deliver the same return as the market index

ETFs hold all the index constituents, or a representative subset of the index constituents.

The advantage of ETFs compared to e.g. investment funds are flexibility, cost-effectiveness and high liquidity. The EFTs are cheap compared to other investment products.

Remember: as with all other types of investments there are risks involved with buying and selling ETFs. Your investment can go up and down and the amount invested can be lost, see [1].

The available data is found in the file `finans1_data.csv` and consists of 96 columns. The first column is a date column and the subsequent 95 columns indicates the weekly returns (i.e. the ratio between the final and initial price for that week minus 1) of 95 ETFs. The column name specifies the name of the ETF.

Descriptive analysis

The first part of the project is to perform a descriptive analysis to examine the data and its quality, in order to get some knowledge about methods and models we can use to analyze the data.

Make a folder on your PC for the project, unzip download and unzip the material from CampusNet into this folder.

First open the data file `finans1_data.csv` for example in RStudio (File>Open File) and see what is in the file. It is seen that the first line contains variable names and that subsequent lines are the actual observations. Variable names and observation values are separated by a ”,”(comma).

You have to make an R-script to be attached to the assignment as an appendix to document the performed analysis. Open the file `finans3.R`, there is a template for the R-script you have to produce.

First set the ”working directory”to where the script and data file is located on your computer

[1]The above sections are written based on the following references: http://www.ishares.com/dk/private/da/literature/brochure/brochure-introducing-ishares-and-etfs-en.pdf and https://falconinvest.dk/hvad-er-en-etf/

```
## Set the working directory

## In RStudio use conveniently the menu "Session->Set Working
## Directory->To Source File Location"
## In R use only "/" for separating in paths (i.e. no backslash)
setwd("Replace with path to where the project files are")
```

Now the data is read into R by

```
## Import the data

## Read the finans1_data.csv file containing the data
wr <- read.table("finans1_data.csv", header=TRUE, sep=",", as.is=TRUE)
```

such that `wr` is a data.frame (table, see the R intro in Chapter 1 of the eNotes) containing the data.

To get an overview of the data first run the following commands

```
## Overview of the data

## Dimension of HE (number of rows and columns)
dim(wr)
## Column names
names(wr)
## The first rows
head(wr)
## The last rows
tail(wr)
## Default summary
summary(wr)
## Another summary function also including the data type
str(wr)
```

a) Make a short description of the data - how many observations does the data set contain, which period is covered, when is the first observation and when is the last observation recorded, how is the data quality e.g. in the form of missing observations? etc.

In the following, we select 4 different ETFs that we will continue to analyze. A description of the selected ETFs listed in the table in Appendix 1.

The Excel file `ETF_dokumentation.xls` contains a description of all the ETFs. The table in Appendix 1 is taken from this file.

b) Examine the empirical distribution/density of each of the 4 EFTs, as a minimum, plot the empirical density and box plots (see eg.Ex1-26 and 1.28).

Further, fill out the following table:

EFT	(Number of obs.)	(Sample mean)	(Sample variance)	(Std. dev.)	(Lower quartile)	(Median)	(Upper quartile)
AGG							
VAW							
IWN							
SPY							

For filling out the table you may run the following commands:

```
## Descriptive analysis of selected variables
## b)
sum(!is.na(wr$AGG))
mean(wr$AGG)
sd(wr$AGG)
## ...

## Alternatively, to run a "function" on the selected columns you
## can use the "apply"-command or wrap it in a for-loop.
## For futher info see ?apply.
```

c) Based on the above analysis, briefly describe the distribution of the weekly returns for each of the 4 ETFs. Are they symmetrical or skewed, and if they are skewed, what kind of skewness. Describe interesting details, including account for extremes / outliers. What values can the weekly returns assume (state perhaps minimum and maximum) and is it as you expect?

Note, that a density is skewed if the probability mass is unevenly distributed (i.e. not symmetrical. If it is a *left-skewed* distribution, the longer tail is located to the left of the median (in general, the mean is located to the left of the median) and similarly for a *right-skewed* distribution the longest tail to the right of the median (in general, the mean is located to the right of the median).

Statistical analysis I

We will now start at the statistical analysis and make statistical inference, hence we will analyze portfolios, formulate models, test hypotheses and determine confidence intervals.

Problem 1 - ETF Portfolio

In relation with the construction of a portfolio of ETFs diversification of risk is a key concept. It's about "not putting all your eggs in one basket". Risk can be measured in several ways - e.g. the standard deviation of the weekly returns. Another used measure of risk is the concept of volatility, which is the standard deviation of the logarithm of the weekly returns (also the same as log (return)).

When you construct a portfolio of ETFs, the covariance between the various ETFs is an essential tool to determine how to allocate your investment between the ETFs (how much you want to invest in the various ETFs).

d) Determine the covariance between the following ETFs: AGG, VAW, IWN, SPY, EWG and EWW. Use the following R code:.

```
## d)
## Determination of the correlation between ETFs
## and determination of portfolio
cov(wr[ ,c("AGG","VAW","IWN","SPY","EWG","EWW")])
```

e) Make a portfolio of two ETFs so that the variance of the portfolio is minimized doing it in following steps:

1. Let P_1 be a random variable that describes the portfolio consisting of EFT'erne: EWG og EWW: $P_1 = \alpha \cdot X_{EWG} + (1 - \alpha) \cdot X_{EWW}$, where X_{EWG} respectively X_{EWW} are random variable that indicates the weekly returns for EWG respectively EWW. α specifies the proportion of the portfolio invested in EWG.

   Define corresponding random variables, describing the following portfolio with two ETFs: (AGG,SPY), (VAW,IWN), (VAW,EWG), (VAW,EWW) og (IWN,EWG). If you like, you can also try additional portfolios of other combinations among ETFs.
2. Determine an expression for $var(P_1)$ (see Remark 2.59 and Theorem 2.60).
3. Determine $var(P_1)$ as a function of α (dvs. $V(\alpha)$), where the values of $var(X_{EWG}$, $var(X_{EWW})$ and $cov(X_{EWG}, X_{EWW})$ (determined in question d)) are inserted.
4. Make a graph of $V(\alpha)$.
5. Determine α_m, that provides minimum variance of P_1 - Consider the monotony conditions for the function. What is the implication of $\alpha_m > 1$ or $\alpha_m < 0$?
6. Make a table of all α_m -values for the investigated portfolio combinations.
7. Make a table of minimum variance for the investigated portfolio combinations.
8. Make a table of the expected weekly returns for the portfolios with minimum variance ($E(P_i)$).
9. Choose the optimal portfolio - arguments for your choice of portfolio.

Problem 2 - Best investment

In this part we will investigate what the best investment is: saving your money under the pillow or investing in one of the 4 selected ETFs.

f) Set up models in which we can assess the weekly returns for each of the 4 ETFs. State the assumptions of the model. Estimate the model parameters and perform model control.

   Carry out a model validation, i.e. examine if the assumptions of the test are fulfilled (see Section 3.1.8).

   The following R code can be used for validation of model:

```r
## Model validation
## f)
## Validation of a model for AGG
qqnorm(wr$AGG, main='Validation of normal distribution assumption for AGG',
       xlab='z-scores', ylab='Weekly returns')
qqline(wr$AGG)
## Do the same for the other ETFs
```

Assess whether assumptions are fulfilled. Remember to also include the Central Limit Theorem in the assessment (Theorem 3.14).

If the normal distribution assumption is not met (important if the distribution is highly skewed and sample size is small) a transformation of data should be considered - typically by the logarithm function. Since this data is financial, it could be considered to investigate the geometric average rather than the arithmetic average. In this assignment you can omit these considerations.

g) Determine a 95% confidence interval for the average weekly return for each of the 4 ETFs, and a 95% confidence interval for the variance parameterfor each of the 4 ETFs; state formulas and insert numbers. Describe the confidence intervals, e.g.: Do they have the same width? If they do not have the same width, what is the reason?

Compare your results of the determination of confidence intervals with the results from the following R code:

```r
## Calculations of the 95% confidence intervals
## g)
## t-quantile for the confidence interval for the mean of AGG,
## since the degrees of freedom for the mean of AGG are 453
qt(0.975, 453)

## Determination of the confidence interval for the mean parameter in a
## normally distributed random sample

## The 95% confidence interval for AGG
t.test(wr$AGG, conf.level=0.95)$conf.int
## Do the same for the other ETFs
```

h) The following questions can be addressed by day 7, so you may wish to wait with the question. In case that you can not achieve that the normal distribution assumption is fulfilled - even after a reasonable transformation, it will still be relevant to examine the empirical distribution and determine both the confidence interval for the average / mean parameters and variance parameters, respectively. Find the mentioned confidence intervals with non-parametric bootstrap and compare the results with the above specific outcomes.

i) Test a hypothesis that the average weekly return do not differ significantly from saving the money under the pillow, similar to the following hypothesis:

$$H_0 : \mu_{\text{AGG}} = 0$$
$$H_1 : \mu_{\text{AGG}} \neq 0$$

It is ok to use the build in R-functions here.

Problem 3

The second analysis we will carry out is to examine if there is a similar weekly return between the four ETFs.

We have yet not learned to analyze this fully, since it requires an analysis of variance (eNote Chapter 8). However, we can make a bit simpler analysis, where we compare the two ETFs that have the lowest and highest average weekly return.

j) Determine the two ETFs with the lowest and highest average weekly return, and do a statistical analysis where the following hypothesis is tested

$$H_0 : \mu_{ETF_{low}} = \mu_{ETF_{high}}$$
$$H_1 : \mu_{ETF_{low}} \neq \mu_{ETF_{high}}$$

Remember to state the model, the significance level α, it is again fine to use the build in R-functions. Explain how p-value/critiucal value is calculated.

Statistical analysis II

In this part of the project, we will derive and investigate a model for return on ETFs. In general, it is anticipated that *high returns are associated with high risk*. Our model will therefore contain a risk measures as an explanatory variable.

In the section "Statistical analysis I", we investigated the weekly return for each ETF in the period May 5, 2006 to May 8, 2015 - a total of 454 weeks. In our model in this section the response variable will be an "average"weekly return.

Let X_t denote the price of an ETF at the end of week t. Such that X_{t-1} denotes then the exchange rate for the week $t-1$.

Within economics (respectively exponential growth) the geometric mean is usually applied to describe the projection factor over time. This and other terms used later are derived here:

The *projection factor* for the t'th week is defined by

$$a_t = 1 + r_t = \frac{X_t}{X_{t-1}} \tag{1}$$

where r_t is the return (the relative return) for the t'th week.

The *total projections factor* of the 454 weeks is

$$a = a_1 \cdot a_2 \cdot \ldots \cdot a_{454} = \prod_{t=1}^{454} a_t \tag{2}$$

and the total relative return over the period is: $r = a - 1$

The *average projection per. period* (here per week) is determined by

$$a_{\text{week}} = \sqrt[454]{a} = \exp\left(\frac{1}{454} \cdot \sum_{t=1}^{454} \log(a_t)\right) \tag{3}$$

which is the *geometric mean*.

The last rewrite is of computational reasons, since a often is a number very close to 1, then higher computational accuracy is achieved, when the logarithmic transformation is used.

The *average relative weekly return* is thus calculated by

$$r_{\text{week}} = a_{\text{week}} - 1 \tag{4}$$

Risk measures

There exists several risk measures - a very simple one is the standard deviation of the exchange rate. In our data, we don't have observations of the price (only the projection), however we can calculate the following risk measures:

- A more common risk measures is the *volatility*, which is the standard deviation of the ratio between the the exchange rate of an ETF in the beginning and the end of a week (i.e. the projection factor a_t)

$$v = 100 \cdot \sqrt{\frac{1}{454} \cdot \sum_{t=1}^{454} (a_t - \bar{a})^2} \tag{5}$$

  where $\bar{a}$ is the average of the projection factor.

- *Value-at-Risk* (VaR) is a another risk measures. VaR is a measure of how much the value of an asset such as an ETF will fall during a given period at a given probability (confidence level) under normal market conditions. The weekly VaR is determined on the basis of a historical distribution of weekly returns as the 5 percent percentile. The assumption is that the future weekly returns distribution is as the historical. VaR thus indicates the highest loss at confidence level of 95 percent and the probability of further losses is 5 percent.

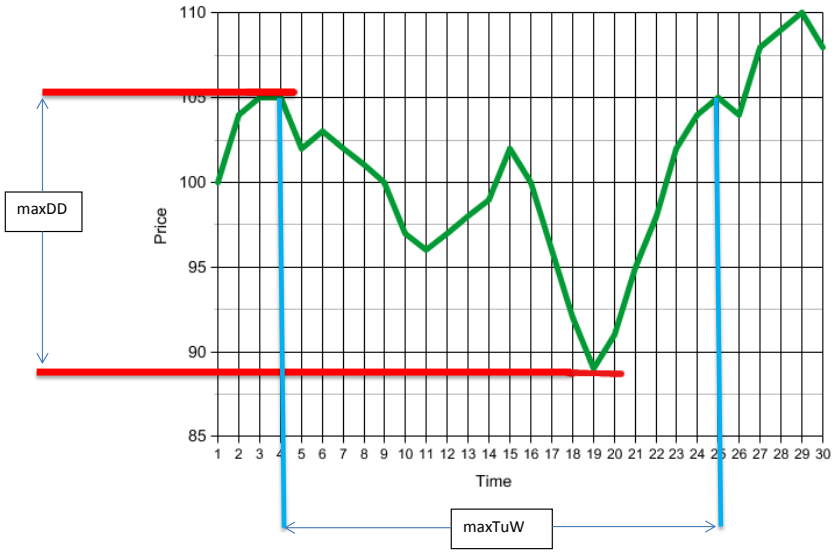
- *Conditional-Value-at-Risk* (CVaR) is another risk measures. CVaR is also referred to as "average value at risk"(AVaR) or "expected tail loss". These terms provide a good understanding of the meaning of risk measures. Where VaR gave the highest loss with a given probability, CVaR indicates the expected / average loss of the 5 percent worst situations.

- MDD or *Maximum Drawdown* is a risk measure that determines the largest possible losses in a given period

$$MDD = \frac{\max(r_t) - \min(r_t)}{\max(r_t)} \tag{6}$$

  MDD is especially used if the asset (ETF) is managed with an active strategy or if the returns from period to period is not independence.

- Note MDD measures nothing about the time between maximum and minimum. One may therefore need the following risk measures "Maximum Time under Water"(In some contexts MaxTuW also called "Maximum Draw Down Duration"). MaxTuW thus indicates the time for regaining historical peak.



Figur .2: The figure illustrates the concepts MDD/maxDD og maxTuW.

Problem 4

The data for this part of the project include observations of 8 variables. Table 1 provides an overview of all variables and their meaning. The data is available in the file `finans2_data.csv` (the description of the individual ETFs can be found in the file `ETF_dokumentation.xls`).

Variable	Meaning	Unit
X	Name of the ETF	
Geo.mean	Geometric mean relative weekly return r_{week}	Pct.
Volatility	The weekly volatility	Pct.
maxDD	Maximum Draw Down	Pct.
maxTuW	Maximum Time under Water	Pct.
VaR	Weekly Value-at-Risk	Pct.
CVaR	Weekly Conditional Value at Risk	Pct.

Table 1: The available variables.

In total there are 95 observations of ETFs
First import data:

```
## Import data finans2_data.csv
etfSum <- read.table("data/finans2_data.csv",header=TRUE, sep=",")
str(etfSum)
```

We want to assess the relation between the numerical variables in Table 1 using their empirical correlations.

```
## Determine the empirical correlation for the selected variables and
## examine the dependencies
cor(etfSum_analyse[,2:7], use="everything", method="pearson")

## First trim the square around the plot. See more on ?par
par(mar=c(3,3,2,1),mgp=c(2,0.7,0))
par(mfrow=c(1,1))
plot(etfSum_analyse$Volatility, etfSum_analyse$CVaR, pch=16, cex=0.7,
     xlab="Volatility [Weekly Pct.]",
     ylab="Conditional Value at Risk [Weekly Pct.]",  cex.lab=0.8,
     main="Relation between Volatility and CVaR", cex.main=0.8)
```

k) Also make scatter plots for Geo.mean vs. maxTuE, Volatility vs. maxDD, and maxTuW vs. Volatility.

   Describe and interpret shortly the empirical correlations. Calculate yourself the empirical correlation between *Geo.mean* and *maxTuW*, state the formulas and insert the values. (See e.g. Definition 1.19 and Remark 1.21 in eNote 1). Compare with the correlation found using R.

```
## k)
cov(etfSum_analyse$Geo.mean, etfSum_analyse$maxTuW)
var(etfSum_analyse$Geo.mean)
var(etfSum_analyse$maxTuW)
```

From the above analysis it should be clear that the various risk measures are highly correlated. Creating a model for the geometric mean return, where all risk measures variables are implemented, will lead to a model with collinearity (This type of model is introduced in eNote 6). The problem with such a model will therefore be that it is difficult to reduce the model in the right way, and it may be difficult to evaluate each individual explanatory variable's importance to the response variable (the geometric average return). We will therefore limit ourselves to a model with only one explanatory risk measure.

l) Formulate a linear regression model with the geometric mean of return *Geo.mean* as the dependent variable (Y_i), and select one of the risk measure variables as the

explanatory variable (x_i). Give reasons for your choice of explanatory variable, as well as stating the assumptions of the model (See eNote 5).

m) Estimate the coefficients in the model, usually called β_0 and β_1, and interpret the estimated values, also estimate the model variance σ^2. Write down the applied equations, also with numerical values inserted in the formulas. What do the estimated values tell about the relation between the geometric mean of return and the explanatory variable?

Verify that you calculated correctly with the following R code:

```
lm1 <- lm(Geo.mean~EXPLANATORY_VARIABLE, etfSum_analyse)
summary(lm1)
```

and interpret some of the remaining values from the R-output.

n) Use the answer of the above to test if there is a correlation between Geo.mean and the chosen risk factor.

Appendix 1

The table shows an overview and description of the 4 selected ETFs.

Excel file `ETF_dokumentation.xls` contains a description of all ETFs. The following table is taken from this file.

ETF	Description
AGG	iShares Core Total US Bond Market ETF, formerly iShares Lehman Aggregate Bond Fund (the Fund) seeks investment results that correspond generally to the price and yield performance of the total United States investment-grade bond market as defined by the Lehman Brothers U.S. Aggregate Index (the Index). The Index measures the performance of the United States investment-grade bond market, which includes investment-grade United States Treasury bonds, government-related bonds, investment-grade corporate bonds, mortgage pass-through securities, commercial mortgage-backed securities and asset-backed securities that are publicly offered for sale in the United States. The securities in the Index must have at least one year remaining to maturity. In addition, the securities must be denominated in United States dollars, and must be fixed rate, non-convertible and taxable. The Index is market capitalization weighted. The Fund uses a representative sampling strategy to track the Index. The Fund's investment advisor is Barclays Global Fund Advisors (BGFA).
VAW	Vanguard Materials ETF (the Fund), formerly known as Vanguard Materials VIPERs, is an exchange-traded share class of Vanguard Materials Index Fund, which employs a passive management or indexing investment approach designed to track the performance of the Morgan Stanley Capital International (MSCI) US Investable Market Materials Index (the Index). The Index is an index of stocks of large, medium and small United States companies in the materials sector, as classified under the Global Industry Classification Standard (GICS). This GICS sector is made up of companies in a range of commodity-related manufacturing industries. Included within this sector are companies that manufacture chemicals, construction materials, glass, paper, forest products and related packaging products, as well as metals, minerals and mining companies, including producers of steel. The Fund attempts to replicate the Index by investing all, or substantially all, of its assets in the stocks that make up the Index, holding each stock in approximately the same proportion as its weighting in the Index. The Fund also may sample its target index by holding stocks that, in the aggregate, are intended to approximate the Index in terms of key characteristics, such as price/earnings ratio, earnings growth and dividend yield.
IWN	iShares Russell 2000 Value Index Fund (the Fund) seeks investment results that correspond generally to the price and yield performance of the Russell 2000 Value Index (the Index). The Index measures the performance of the small-capitalization value sector of the United States equity market. It is a subset of the Russell 2000 Index. The Index is a capitalization-weighted index and consists of those companies or portion of a company, with lower price-to-book ratios and lower forecasted growth within the Russell 2000 Index. The Index represents approximately 50% of the total market capitalization of the Russell 2000 Index. The Fund invests in a representative sample of securities included in the Index that collectively has an investment profile similar to the Index. iShares Russell 2000 Value Index Fund's investment advisor is Barclays Global Fund Advisors.
SPY	SPDR Trust, Series 1 (the Trust) is a unit investment trust. The Trust is an exchange-traded fund created to provide investors with the opportunity to purchase a security representing a proportionate undivided interest in a portfolio of securities consisting of substantially all of the common stocks, in substantially the same weighting, which comprise the Standard and Poor's 500 Composite Price Index (the SP Index). Each unit of fractional undivided interest in the Trust is referred to as a Standard and Poor's Depositary Receipt (SPDR). The Trust utilizes a full replication approach. With this approach, all 500 securities of the Index are owned by the Trust in their approximate market capitalization weight.
EWG	iShares MSCI Germany Index Fund (the Fund) seeks to provide investment results that correspond generally to the price and yield performance of publicly traded securities in the aggregate in the German market, as measured by the MSCI Germany Index (the Index). The Index seeks to measure the performance of the German equity market. The Index is a capitalization-weighted index that aims to capture 85% of the (publicly available) total market capitalization. Component companies are adjusted for available float and must meet objective criteria for inclusion in the Index. The Index is reviewed quarterly.The Fund invests in a representative sample of securities included in the Index that collectively has an investment profile similar to the Index. The Fund's investment advisor is Barclays Global Fund Advisors.
EWW	iShares MSCI Mexico Index Fund (the Fund) seeks to provide investment results that correspond generally to the price and yield performance of publicly traded securities in the aggregate in the Mexican market, as measured by the MSCI Mexico Index (the Index). The Index seeks to measure the performance of the Mexican equity market. The Index is a capitalization-weighted index that aims to capture 85% of the (publicly available) total market capitalization. Component companies are adjusted for available float and must meet objective criteria for inclusion in the Index. The Index is reviewed quarterly.The Fund invests in a representative sample of securities included in the Index that collectively has an investment profile similar to the Index. The Fund's investment advisor is Barclays Global Fund Advisors.