The evolution and gender/age differences of the Swedish households ownerships of stocks listed on the Swedish market 1999-2024

By: Kevin Nilsson

2025-01-09

Introduction

In this report a dataset about: Swedish households ownership of stocks in companies listed on the Swedish stock market by gender and age (1999-2024), has been used. This dataset is collected from SCB (Statistics Sweden), and focus in this report is to provide associations between different variables, and visualize the differences in proportions of the Swedish households investment in Swedish listed stocks between age-groups and genders.

- Are there any significant differences between age-groups and genders in the context of investments in Swedish listed stocks?
- How does the trend develop through the years?
- Have Swedish listed stocks been a lucrative investment between 1999 and 2024?

Different visualizations has been made, using simple bar charts, stacked and clustered bar charts. Correlation analysis is used to see the strength of associations and relationships between variables. Finally, a best model was developed and visualized in a scatter plot which gave information about the most significant factors impacting total investment in Swedish stocks.

```
#Loading libraries
library(ggplot2)
library(dplyr)
library(tidyverse)
library(purrr)
library(car)
```

```
#Loading the dataset
portfolio <- read.csv("C:/Users/Kevin/Desktop/portfolio1.csv", quote="'", header=FALSE)

#Cleaning/Reshaping the dataset
#Switching so that the first row become headers, then remove the first row
colnames(portfolio) <- portfolio[1, ]
portfolio <- portfolio[-1, ]

#Including characters å,ä,ö
names(portfolio) <- iconv(names(portfolio), from="latin1", to="UTF-8") #column names
portfolio <- portfolio %>%
  mutate(across(where(is.character),~iconv(., from="latin1", to="UTF-8")))
```

```r
#Removing characters: M(0-9) from column names
names(portfolio) <- gsub('"|M[0-9]+', '', names(portfolio))

#Removing " marks from the columns kön and ålder
portfolio$kön <- gsub('"', '', portfolio$kön)
portfolio$ålder <- gsub('"', '', portfolio$ålder)

#Renaming "män" to "men" and "kvinnor" to "women" in the kön-column
portfolio$kön <- gsub('män', 'men', portfolio$kön)
portfolio$kön <- gsub('kvinnor', 'women', portfolio$kön)

#Renaming "år" to "years" in the 'ålder' column
portfolio$ålder <- gsub('år', 'years', portfolio$ålder)

#Renaming columns ålder and kön into age and gender
colnames(portfolio) <- gsub('ålder', 'age', colnames(portfolio))
colnames(portfolio) <- gsub('kön', 'gender', colnames(portfolio))

#Reshaping from wide to long format
longportfolio <- portfolio %>%
  pivot_longer(cols=starts_with("20") | starts_with("19"),
               names_to = "Year", #naming the new column
               values_to = "Households")

#Removing " marks from the column Households
longportfolio$Households <- gsub('"', '', longportfolio$Households)

#Checking for missing values
missing_values_longportfolio <- sum(is.na(longportfolio))

#Glimpse of the dataset
glimpse(longportfolio)
```

```
## Rows: 500
## Columns: 4
## $ gender     <chr> "women", "women", "women", "women", "women", "women", "wome~
## $ age        <chr> "18-24 years", "18-24 years", "18-24 years", "18-24 years",~
## $ Year       <chr> "1999", "2000", "2000", "2001", "2001", "2002", "2002", "20~
## $ Households <chr> "7373", "7583", "6744", "5969", "5956", "3997", "3826", "38~
```

Describing the table above:

After loading the dataset, doing some data cleaning and preparations to make it easier to work with, a glimpse was taken (as shown in the table above). Here we get the information that the dataset consists of 500 rows, 4 columns, and the variables are all characters. I will change "age" and "gender" into factors because they all include different categories. The variables "Households" and "Year" will be changed into numeric kind, this is done for easier analysis purpose. Also, checked for missing values, which equals = 0.

```r
#Factor transformations
longportfolio$gender <- as.factor(longportfolio$gender)
longportfolio$age <- as.factor(longportfolio$age)

#Numeric transformations
```

```r
longportfolio$Year <- as.numeric(longportfolio$Year)
longportfolio$Households <- as.numeric(longportfolio$Households)

#Glimpse again to see that the transformations were successful
glimpse(longportfolio)
```

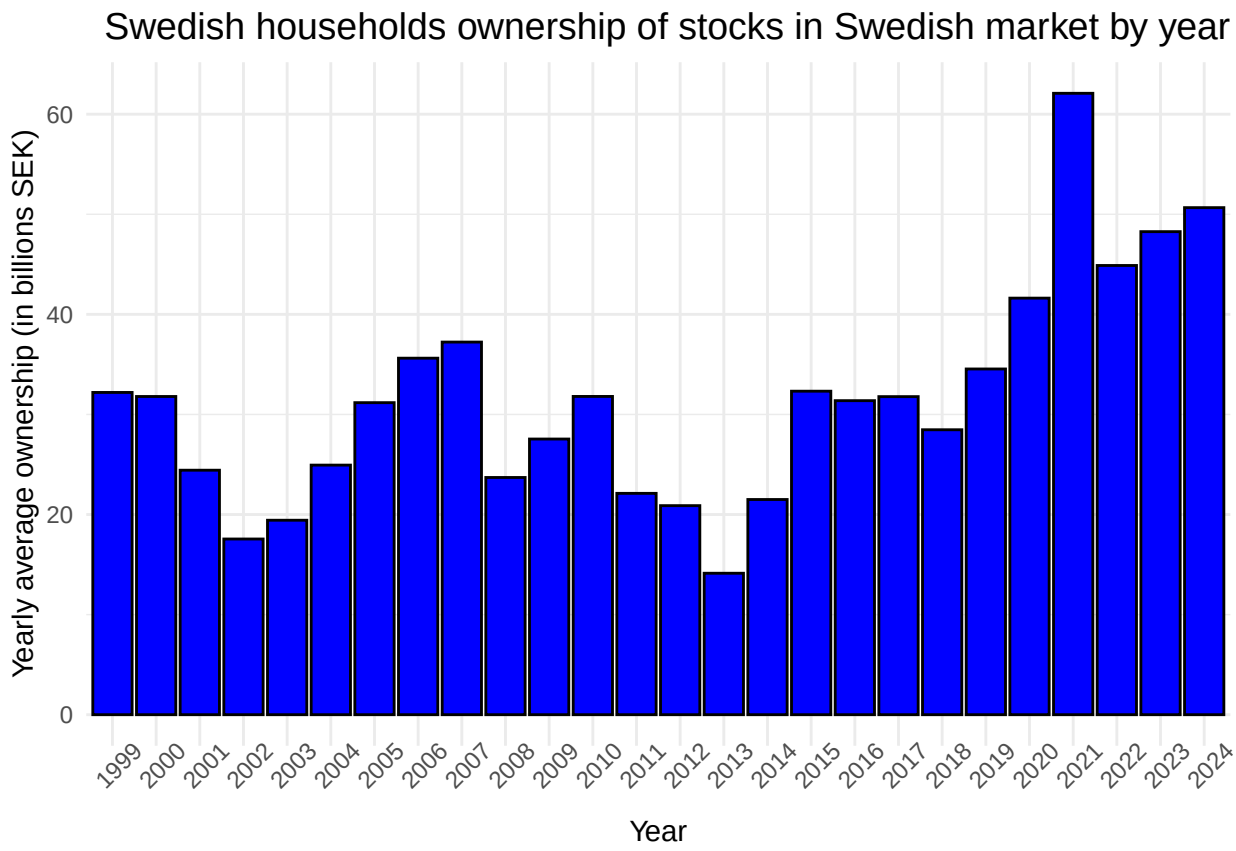
```
## Rows: 500
## Columns: 4
## $ gender     <fct> women, women, women, women, women, women, women, women, wom~
## $ age        <fct> 18-24 years, 18-24 years, 18-24 years, 18-24 years, 18-24 y~
## $ Year       <dbl> 1999, 2000, 2000, 2001, 2001, 2002, 2002, 2003, 2003, 2004,~
## $ Households <dbl> 7373, 7583, 6744, 5969, 5956, 3997, 3826, 3870, 4755, 3665,~
```

Commenting on the table above:

In the table above we see that the factor and numeric transformations were successfully executed. Now some visualizations of variables and associations will be made.

```r
#Summarizing to one value per year using the mean because I want to specifically study-
#time-trends by year
port_data <- longportfolio %>%
  group_by(Year) %>%
  summarize(Average_Households=mean(Households))

#Bar chart showing the trend by year of Swedish households ownership of shares (in SEK)-
#listed on the Swedish stock market
ggplot(port_data, aes(x=factor(Year), y=Average_Households/1e3))+
  geom_bar(stat="identity", position="dodge", color="black", fill="blue")+
  theme_minimal()+
    labs(
      title="Swedish households ownership of stocks in Swedish market by year",
    x="Year",
    y="Yearly average ownership (in billions SEK)")+
  theme(
    plot.title=element_text(size=14, hjust=1),
    axis.text.x=element_text(angle=45, hjust=0.5))
```



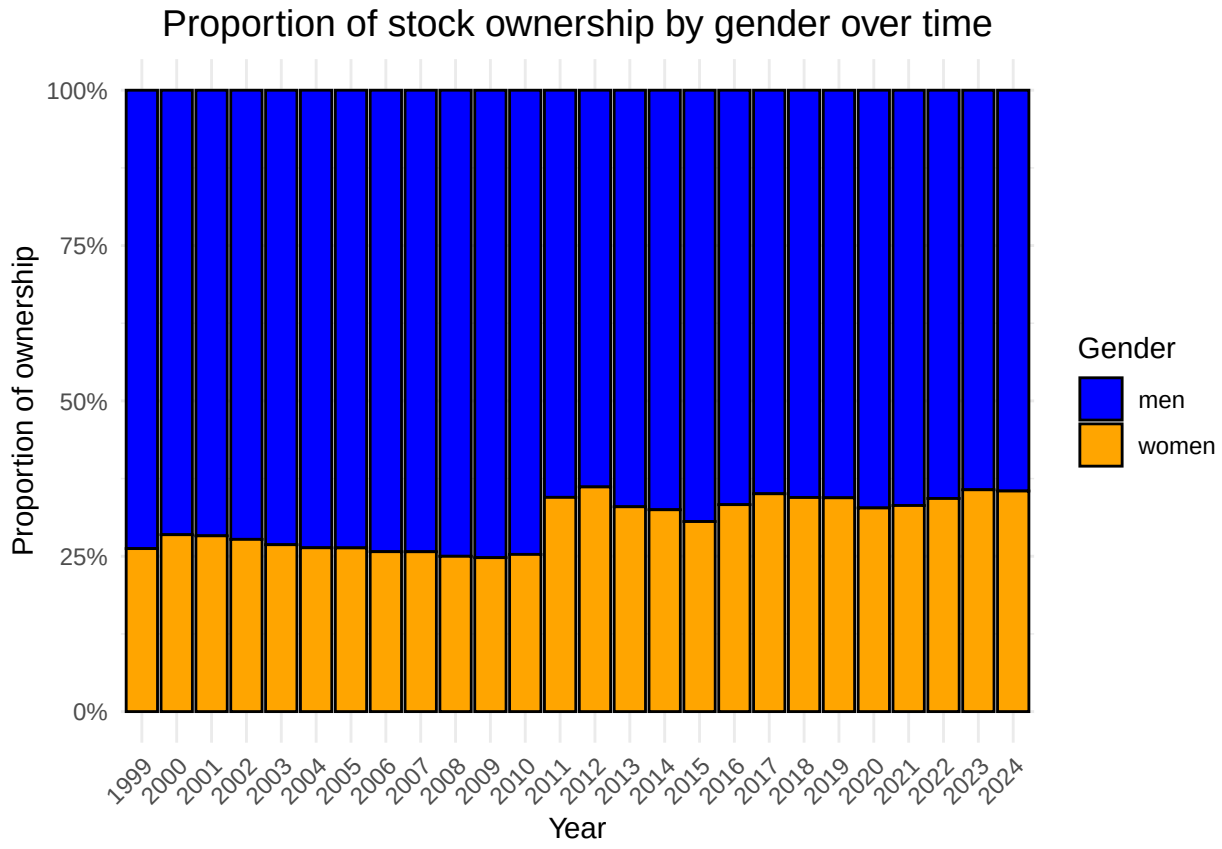
By analyzing the bar chart from above, which is showing Year on the x-axis and Yearly average ownership (in billions SEK) on the y-axis. Note that data is collected 2x for each year, but here the mean value of these two observations per year is shown. Averaged over all age-groups and genders. So interpretations can be read like this: Each bar represents the average stock ownership (in billions SEK) across all gender and age groups for each year. There are two notable weaker spots in the plot, shown in 2002 and 2013. These could be explained by the dot-com bubble in 2000 which lead to a recession, hence the Swedish economy slowed down a bit, but to come back stronger again and keep increasing between 2003-2007. In 2008 there was a global financial crisis which affected the market a lot. This is shown by the steady decrease in the yearly average ownership between 2008 and 2013.

After this there has been a steadily increase in the Swedish stock market. Even though the covid-19 pandemic hit hard, but it had a positive effect on new investors, as the invest-interest grew a lot due to pandemic restrictions, remote working, investing-influences on social media etc. Hence, we can see a significant increase for the year 2021. For 2021 the Swedish households owner-ship in stocks listed on the Swedish stock market reaches a level equal to slightly above 60 billions SEK when averaging over gender, age-groups and years.

```
#Calculate proportions of ownership by gender and year
port_data_gender_prop <- longportfolio %>%
  group_by(Year, gender) %>%
  summarize(Tot_Households=sum(Households, na.rm=TRUE)) %>%
  mutate(Proportion=Tot_Households/sum(Tot_Households)) %>%
  ungroup()

#Stacked bar chart showing proportions
ggplot(port_data_gender_prop, aes(x=factor(Year), y=Proportion, fill=gender)) +
  geom_bar(stat="identity", color="black") +
  scale_y_continuous(labels=scales::percent_format())+
```

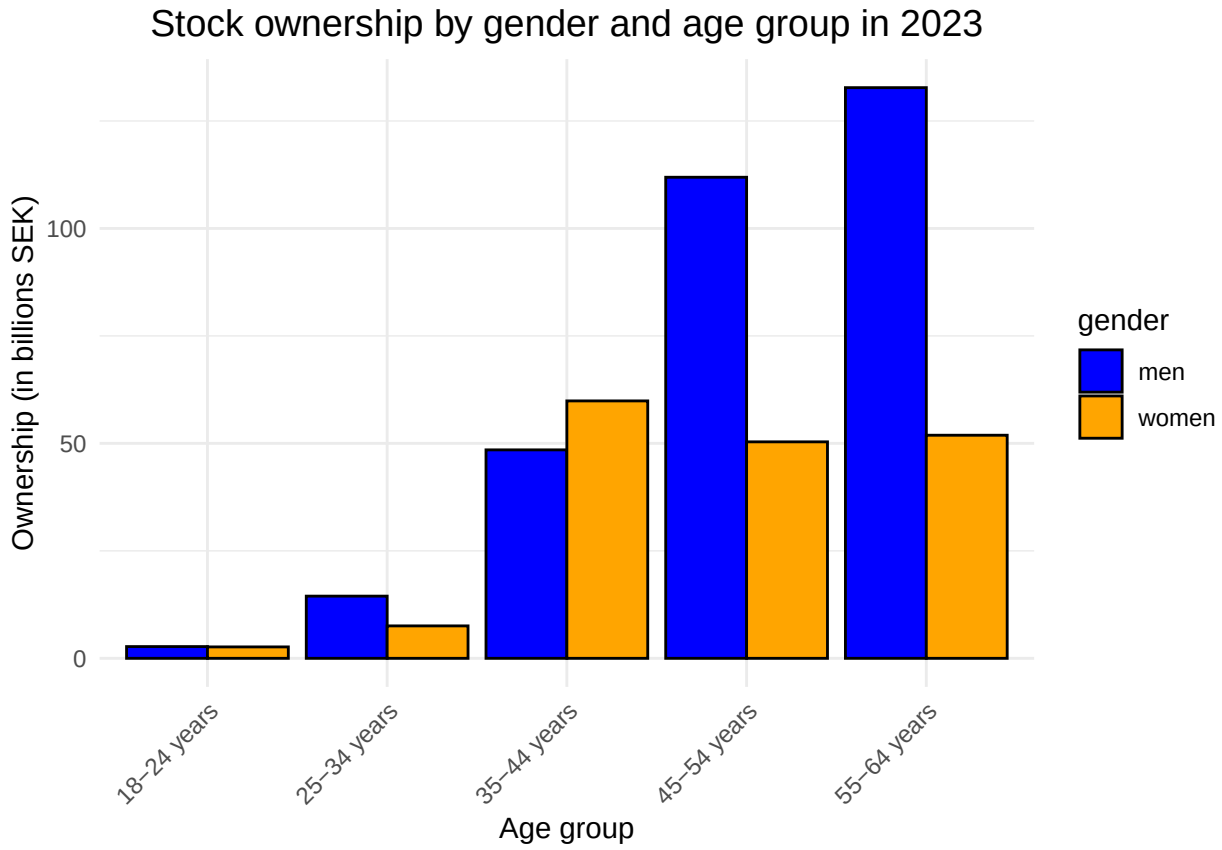
```
  scale_fill_manual(values=c("men"="blue", "women"="orange"))+
  theme_minimal()+
  labs(
    title="Proportion of stock ownership by gender over time",
    x="Year",
    y="Proportion of ownership",
    fill="Gender")+
  theme(
    plot.title=element_text(size=14, hjust=0.5),
    axis.text.x=element_text(angle=45, hjust=1))
```



This stacked bar chart from above provides information about the proportions (gender-wise) of Swedish households ownership in stocks that is listed on the Swedish stock market. Men-proportion is shown in the color blue, and women-proportion is shown in the color orange. Here its visual that men stands for the largest proportion of investment in monetary value for all the years (1999-2024), around 70-75%. So, conclusion can be made that in the Swedish households, the men owns the largest proportion of stocks listed on the Swedish stock market in comparison to the women. Worth to mention also, is that the proportional differences of investments decrease a bit in the later year as women stands for approx. 35% and men for approx. 65% from 2011 and forward. Note that in this plot we do not see differences in age-groups, this will be shown in the next plot specifically for the year 2023.

```
#Average the stock values for 2023 by gender and age group
data_2023_avg <- longportfolio %>%
  filter(Year==2023) %>%
  group_by(gender, age) %>%
  summarize(Stock_val=mean(Households))
```

```r
#Bar chart showing stock ownership by age and gender for 2024
ggplot(data_2023_avg, aes(x=factor(age), y=Stock_val/1e3, fill=gender))+
  geom_bar(stat="identity", position="dodge", color="black")+
  scale_fill_manual(values=c("men"="blue", "women"="orange"))+
  theme_minimal()+
  labs(
    title="Stock ownership by gender and age group in 2023",
    x="Age group",
    y="Ownership (in billions SEK)")+
  theme(
    plot.title=element_text(size=14, hjust=0.5),
    axis.text.x=element_text(angle=45, hjust=1))
```



Above we can see a clustered bar chart with data from the year 2023, which shows different age-groups on the x-axis, and genders in different colored bars (men=blue, women=orange). On the y-axis we have Swedish households ownership in stocks listed in the Swedish stock market, in billions SEK.

For the age-group 18-24 year-olds, the proportions looks similar for both genders equally approx. 2.7 billions SEK, then for 25-34 year olds we see that the men has more money invested in Swedish stocks than that of the women (approx. 14.5 billion SEK for men, and 7.5 billion SEK for women). But this changes for the age-group 35-44 year-olds, here the women has a clear larger amount of money invested than that of the men (approx. 48.5 billion SEK for men and 60 billion SEK for women).

When reaching the age of 45+ we see that the men has a significantly higher amount of money invested in stocks listed in the Swedish market, as a matter of fact more than double the amount. Also, the monetary value invested in Swedish stocks for men has increased alot, from 48.5 billion for 35-44 year olds, to 112

billion 45-54, and even higher at 133 billion SEK at 55+. As for the women there is actually a decrease in monetary invested value after the age of 44. There can be many reasons of why, but since the main focus of this report is only to display the differences I wont go deeper into this.

Below here, fitting of regression models is made, showing summaries for the different models and comparing them through adj. R-squared and AIC values in order to make conclusion of the best model. This is done in order to trying to predict the variable "Household" (Total monetary value invested by Swedish households into Swedish listed stocks) and which variables that are most important in doing so.

```
#Fitting linear models
model1 <- lm(Households~gender, data=longportfolio)
model2 <- lm(Households~gender+Year, data=longportfolio)
model3 <- lm(Households~gender+Year+age, data=longportfolio)

#Summary model 1
summary(model1)
```

```
##
## Call:
## lm(formula = Households ~ gender, data = longportfolio)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41245 -17219  -8100  14318 142574
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    43215       2018  21.414  < 2e-16 ***
## genderwomen   -23979       2854  -8.402 4.63e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31910 on 498 degrees of freedom
## Multiple R-squared:  0.1241, Adjusted R-squared:  0.1224
## F-statistic: 70.59 on 1 and 498 DF,  p-value: 4.629e-16
```

```
#Summary model 2
summary(model2)
```

```
##
## Call:
## lm(formula = Households ~ gender + Year, data = longportfolio)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51134 -19620  -5172  13529 146454
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1691181.7   389821.9  -4.338 1.74e-05 ***
## genderwomen   -23978.8     2801.6  -8.559  < 2e-16 ***
## Year             862.2      193.8   4.449 1.06e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31320 on 497 degrees of freedom
## Multiple R-squared:  0.1577, Adjusted R-squared:  0.1543
## F-statistic: 46.52 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
#Summary model 3
summary(model3)
```

```
##
## Call:
## lm(formula = Households ~ gender + Year + age, data = longportfolio)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47825 -14501  -3370   9071 110840
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1718788.7   270698.4  -6.349 4.91e-10 ***
## genderwomen      -23978.8     1945.5 -12.326  < 2e-16 ***
## Year                862.2      134.6   6.407 3.46e-10 ***
## age25-34 years     9553.4     3076.0   3.106  0.00201 **
## age35-44 years    25100.6     3076.0   8.160 2.81e-15 ***
## age45-54 years    39657.6     3076.0  12.892  < 2e-16 ***
## age55-64 years    63723.1     3076.0  20.716  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21750 on 493 degrees of freedom
## Multiple R-squared:  0.5971, Adjusted R-squared:  0.5922
## F-statistic: 121.8 on 6 and 493 DF,  p-value: < 2.2e-16
```

Model1: Households = 43215 - 23979genderwomen

In this model the independent variable is statistically significant when testing on all levels of significance. The model gets an adj. R-squared = 0.1224 which is pretty low, so there are a lot of room for improvement when trying to predict the total amount of monetary value in Swedish listed stocks from the Swedish households.

Model2: Households = -1691181.7 - 23978.8genderwomen + 862.2Year

By adding the variable Year to the model, our adj. R-squared value increases to 0.1543, so now the independent variable explain 15.43% of the variability in the Households variable. Also here, all independent variables are statistically significant on all level of significance. Yet, still room for alot of improvement.

Model3: Households = -1718788.7 - 23978.8genderwomen + 862.2Year + 9553.4age25-34 + 25100.6age35-44 + 39657.6age45-54 + 63723.1age55-64

So, by adding the variable age to the model (which is a factor variable with different age categories) this boosts the adj. R-squared all the way up to 0.5922, so this helps a lot in explaining the variability in the dependent variable. Now 59.22% of the variability in Swedish households total monetary value in the Swedish listed stocks are explained by year, age-groups and gender. I will also check the AIC values of the models in order to make a final conclusion of which model that is the best, although right now it certainly looks like the Model 3 is the best. Hence, conclusion can be made that age is the most important/most impact able variable when trying to predict total monetary value of Swedish households total monetary value in Swedish listed stocks. This is proven by the significant increase in adj- R-squared.

-
-
-

```r
#Calculating the AIC values for each model
AIC(model1, model2, model3)
```

```
##        df      AIC
## model1  3 11793.58
## model2  4 11776.05
## model3  8 11415.30
```

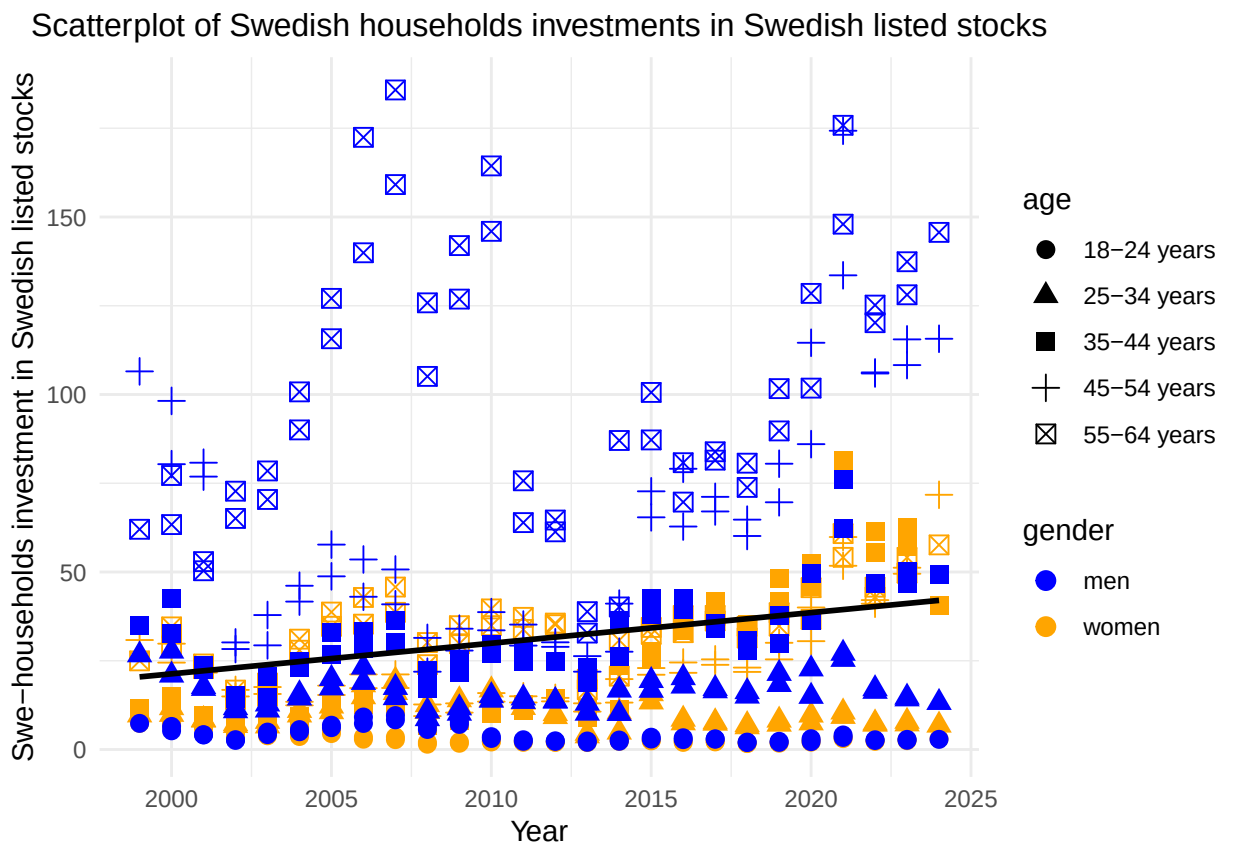
Here, in the table showing AIC-values of the different models above, these are collected in order to reduce over-fitting issues when trying to predict the dependent variable. We clearly see that model3 has the lowest AIC value. Hence, conclusion can be made that the Model3 is the best model in explaining the Swedish households investments in the Swedish listed stocks. I will also check for multicollinearity, because one important assumption in the linear regression is that there is no multicollinearity amongs the independent variables of the model. VIF-values are used to evaluate this.

```r
#Computing VIF to evaluate multicollinearity
vif_results <- vif(model3)
print(vif_results)
```

```
##        GVIF Df GVIF^(1/(2*Df))
## gender    1  1               1
## Year      1  1               1
## age       1  4               1
```

When checking the VIF values of the coefficients, we see that all has values = 1, this tells us that there is no sign of multicollinearity in the model. Hence, no independent variables in this model are highly correlated with each other.

Now, lets visualize the best model in a scatter plot.



The scatter plot above shows monetary value on the y-axis in billions SEK, and years on the x-axis. Men and women are divided into different colors, men=blue and women=orange. There is also a best-fitted line shown in black which tells us that the Swedish households investmed monetary value in Swedish listed stocks are increasing by the years because of its positive slope. The plot includes data collected for each age-groups 2 times/year, except the years 1999 and 2024 where data has only been collected 1 time. We can clearly see that people in the Swedish households of the lower ages are having less money invested in Swedish listed stocks (shown in circles and squares), and the amount of money invested increases significantly for people above the age of 45 (shown in + signs, and squares with an x inside of it). Also seen here is that the women has more money invested in Swedish stocks for the age group 35-44 year olds, in the later years. For the rest of the age-groups and years men seems to be the category with larger amount of money invested here, which can be due to lots of reasons as written earlier in the report and I wont go deeper into this because this report is only to display the differences in this specific case.

Summary

So, to summarize this report, different visualizations has been made. This showed an increasing trend in investing in Swedish listed stocks through the years as for the Swedish households. We could clearly see that the total monetary value has increased on the long term, which tells us that investing in stocks listed on the Swedish market has been a good investment between the years 1999-2024, with a couple of tough years as a cause of economic crisis, but in the long run the stock values has become even larger, with new all-time highs.

When analyzing the differences in stock-ownership by gender, we could see that through all the years 1999-2024 men has the significantly larger proportion of the total amount invested by the Swedish households, as a matter of fact men stands for 75% of the proportion for most of the years. However, this changes a bit in

the later years, as from 2011 and forward the proportions are 35% for women and 65% for men in terms of monetary value invested in Swedish listed stocks.

By taking a more detailed look at the year 2023, to see gender and age-group differences, we saw that 18-24 year olds are similar in monetary amount invested into Swedish listed stocks. 25-34 year olds, the men has the leading edge. 35-44 year olds, women are taking the lead. And from 45+ the men has a significant larger amount invested into Swedish listed stocks.

After this, I fitted 3 regression models and examined the different adj. R-squared, and looked at the AIC values to draw conclusion about the best model. Which resulted in model 3, including all variables and an adj. R-squared = 0.5922.

Model3: Households = -1718788.7 - 23978.8genderwomen + 862.2Year + 9553.4age25to34 + 25100.6age35to44 + 39657.6age45to54 + 63723.1age55to64

The findings of this report underscore the importance of demographic and temporal trends in understanding investment behaviors and their implications for market strategies.

Also, once again this data-analysis is based on data collected from SCB (Statistics Sweden).