

# Analysis of Income, Geographical Location, and Voting Behaviour in Finland

Quy Anh Nguyen

September 21, 2024

## 1 Introduction

### 1.1 The research question

This report investigates how citizens' voting behaviour is potentially influenced by their income and geographical location in Finland.

### 1.2 The materials used

Two datasets are used in this investigation:

- tulot2017.csv: This dataset contains information about how much people earn from wages and investments, how much of those earnings are subject to taxation, and how much people get to keep after taxation. This information is recorded on a per municipality basis in 2017 by Statistics Finland.
- ek2023.csv: This dataset contains information about how much support each political party receives in each municipality, as measured in percentage of total votes in the corresponding municipality. The data is collected in 2019 by Statistics Finland.

### 1.3 How the research is carried out

- First, we explore the datasets to get a preliminary understanding of what they contain.
- Second, we perform some visualisations to identify potential trends and correlations.
- Third, based on the trends identified in the previous step, we formulate hypotheses and perform relevant tests and/or analyses for these hypotheses. We pre-establish that the significance level for any statistical test is 0.01.
- Finally, we summarise and report our findings.

## 2 Data Exploration and Environment Setup

Let us first have a look at our datasets in order to understand them. The first dataset, tulot2017.csv, has the following structure:

```
> income <- read.csv("tulot2017.csv", encoding = "latin1")
> str(income)
'data.frame': 312 obs. of 10 variables:
 $ Alue      : chr  "KOKO MAA" "Akaa" "Alajärvi" "Alavieska" ...
 $ Tulonsaajia : int  4634226 13751 8156 2140 9908 7147 3945 3183 450 879 ...
 $ Tulot      : int  29962 27702 23775 24389 23917 27254 33659 28693 31794 28671 ...
 $ Mediaanitulot : int  24433 23848 19594 20505 20125 22062 26476 26134 25075 26057 ...
 $ Ansiotulot   : int  27801 26421 21962 22459 21981 24447 29008 27235 27725 25954 ...
 $ Pääomatulot  : int  2161 1281 1813 1930 1937 2807 4652 1458 4068 2717 ...
 $ Verot        : int  6453 5657 4463 4605 4436 5561 7461 5889 5949 4997 ...
 $ Valtionvero   : int  1796 1131 913 966 953 1487 2533 1135 2285 1452 ...
 $ Kunnallisvero : int  3997 3867 2985 3038 2911 3454 4183 4073 3054 2985 ...
 $ Tulot_miinus_verot : int  23509 22045 19312 19784 19482 21693 26198 22804 25844 23674 ...
```

The meanings of the fields are as follows:

- Alue: Consists of the different municipalities of Finland.
- Tulonsaajia: Number of taxable income recipients. Essentially, this field records how many people are earning money and are subject to taxation in each of the above municipalities.
- Tulot: Average taxable income, in euros. This field records the average per capita income of people in each municipality, prior to taxation.
- Mediaanitulot: Median taxable income, in euros. This field records the median income of people in each municipality, prior to taxation.
- Ansiotulot: Average earned income, in euros. This field records the average per capita income from work, i.e. wage labour, of people in each municipality, prior to taxation. In other words, this is the average amount of money people are paid by working as an employee for an employer.
- Pääomatulot: Average investment income, in euros. This field records the average per capita income from capital investments in each municipality, prior to taxation. Investment income is income not earned by working, i.e. performing wage labour, but by investing money/capital in stocks, bonds, index funds, real estates, and so on.
- Verot: Average total of all taxes, in euros. This field records the total amount of taxes people pay on average in each of the municipality.
- Valtionvero: Average state tax, in euros. This field records the amount of state tax people pay on average in each of the municipalities. State tax contributes to the total amount of taxes above.
- Kunnallisvero: Average municipal tax, in euros. This field records the amount of municipal tax people pay on average in each of the municipalities. Municipal tax contributes to the total amount of taxes above.
- Tulot\_miinus\_verot: Average income after tax, in euros. This field records how much money people actually get to keep on average in each of the municipalities after all the different types of taxes.

The second dataset, ek2023.csv, has the following structure:

```
> finland <- read.csv("ek2023.csv", encoding = "latin1")
> str(finland)
'data.frame': 293 obs. of 25 variables:
 $ Alue      : chr  "Helsinki" "Askola" "Espoo" "Hanko" ...
 $ SDP       : num  20.9 16.4 16.9 28.5 23.7 25.2 11 24.7 17.6 8.1 ...
 $ PS        : num  11.3 30.8 12.5 16.6 20 24.9 10.7 21.4 25.7 7.9 ...
 $ KOK       : num  26.4 18.8 36.8 10.8 24.5 21.2 11 23.3 15.6 39.3 ...
 $ KESK      : num  1.6 15.1 2.7 1.5 3.7 5.8 0.9 6 7.9 1.7 ...
 $ VIHR      : num  15.3 2.7 10.8 3.4 8.1 5.6 3.3 7.3 3.5 5.7 ...
 $ VAS       : num  11.8 3.2 3.6 4.1 6 5.1 1.9 5.2 19.2 1.7 ...
 $ RKP       : num  5.1 2.9 7.5 27.6 2.6 1 56 1.2 1 29.8 ...
 $ KD        : num  1.9 2.5 3.1 2.8 3.8 3.8 1.9 4.2 3 2.5 ...
 $ PIR       : num  0.3 0.1 0.1 0.1 0.2 0.1 0 0.1 0 0.1 ...
 $ FP        : num  0.1 0.1 0.1 0 0.1 0.1 0.1 0.1 0 0 ...
 $ LIBE      : num  0.9 0.2 1.3 0.3 0.8 0.5 0.3 0.7 0.3 0.6 ...
 $ SKP       : num  0.2 0 0.1 0.1 0.1 0.1 0 0.1 0.1 0 ...
 $ EOP       : num  0.2 0.2 0.1 0.2 0.1 0.1 0.3 0.1 0.1 0 ...
 $ SKE       : num  0.1 0 0 0 0 0 0.1 0 0 0 ...
 $ AP        : num  0.2 0 0 0 0 0 0 0 0 0 ...
 $ KaL       : num  0 0.1 0 0 0 0 0 0 0 0 ...
 $ KL        : num  0.2 0 0 0 0 0 0.1 0.1 0 0 ...
 $ KRIP      : num  0.2 0.2 0.1 0.2 0.2 0.2 0.1 0.2 0.2 0.1 ...
 $ LIIKE     : num  2.3 4.6 2.9 2.8 4.3 4.4 1.3 4 3.8 1.8 ...
 $ SML       : num  0 0.4 0.2 0.2 0.3 0.4 0.1 0.3 0.3 0.2 ...
 $ VKK       : num  0.3 0.4 0.3 0.3 0.4 0.4 0.1 0.3 0.3 0.1 ...
```

```
$ VL      : num  0.8 1.3 0.6 0.4 0.9 1.1 0.7 0.8 1.3 0.4 ...
$ Muut    : num  0 0 0 0.1 0.1 0.1 0.1 0.1 0.1 0 ...
$ Vaalipiiri: chr  "Helsinki" "Uusimaa" "Uusimaa" "Uusimaa" ...
```

Again, the meanings of the fields are as follows:

- Alue: Consists of the different [municipalities](#) of Finland.
- SDP, PS, KOK, ..., Muut: All these fields record how much support each party receives, as measured in percentage of votes in each municipality.
- Vaalipiiri: Consists of the [electoral districts](#) of Finland. One electoral district may comprise multiple municipalities.

To make it easier to deal with these two datasets, we are going to merge them into one data frame called `df`. This is done by connecting the two data sets via the `Alue` field:

```
> df <- merge(finland, income, by = "Alue")
> colnames(df)
 [1] "Alue"          "SDP"           "PS"
 [4] "KOK"           "KESK"          "VIHR"
 [7] "VAS"           "RKP"           "KD"
[10] "PIR"           "FP"            "LIBE"
[13] "SKP"           "EOP"           "SKE"
[16] "AP"            "KaL"           "KL"
[19] "KRIP"          "LIIKE"         "SML"
[22] "VKK"           "VL"            "Muut"
[25] "Vaalipiiri"    "Tulonsaajia"   "Tulot"
[28] "Mediaanitulot" "Ansiotulot"    "Pääomatulot"
[31] "Verot"         "Valtionvero"   "Kunnallisvero"
[34] "Tulot_miinus_verot"
```

The `df` data frame will now contain all the fields from the two data sets, which is more convenient for us. We're also going to be using the `tidyverse` and `ggplot2` libraries:

```
library(tidyverse)
library(ggplot2)
```

### 3 Trend Identification

First, let us see if there is any correlation between income and support for a certain party. For instance, we can plot average taxable income against support for the Social Democratic Party as follows:

```
ggplot(data = df, mapping = aes(x = Tulot, y = SDP)) +
  geom_point() +
  labs(
    x = "Average Taxable Income (euros)",
    y = "Support for the SDP (%)"
  )
```

Running the above code generates the scatterplot in Fig. 1, in which each point represents a municipality. As the figure demonstrates, the percentage of support for the Social Democratic Party mostly falls in the 0-30% range. There also seems to be a very slight positive correlation. We can generalise the above code into the following function:

```
plot_income_against_support <- function(income_type, party) {
  income_type_english <- ""
  if (income_type == "Tulot") {
    income_type_english <- "Average Taxable Income"
  }
  else if (income_type == "Mediaanitulot") {
    income_type_english <- "Median Taxable Income"
  }
}
```

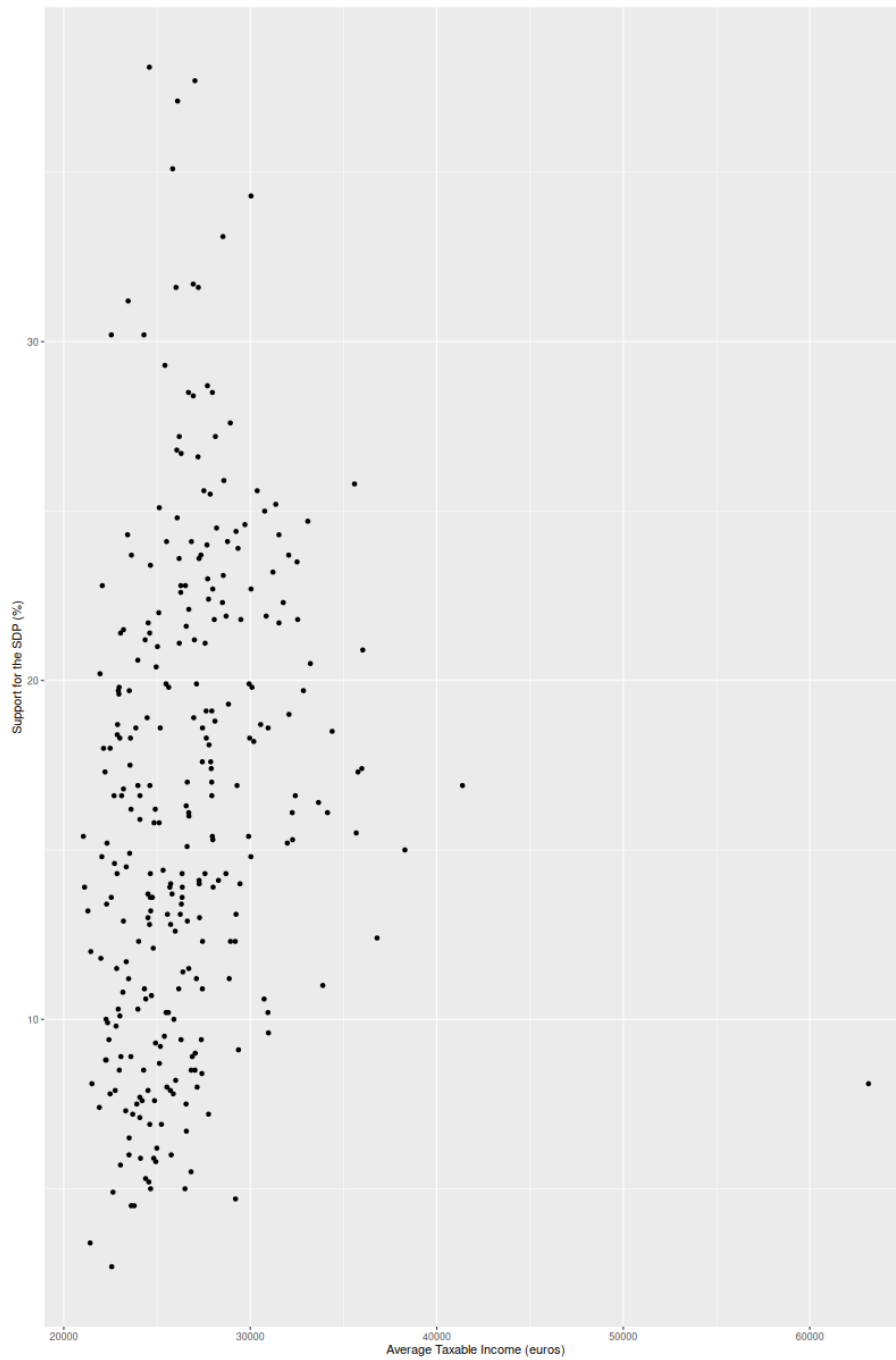


Figure 1: Support for the Social Democratic Party against Average Taxable Income

```

}
else if (income_type == "Ansiotulot") {
  income_type_english <- "Average Earned Income"
}
else if (income_type == "Pääomatulot") {
  income_type_english <- "Average Investment Income"
}
else if (income_type == "Tulot_miinus_verot") {
  income_type_english <- "Average Income after Tax"
}
ggplot(data = df, mapping = aes(x = .data[[income_type]], y = .data[[party]])) +
  geom_point() +
  labs(
    title = paste("Support for", party, "against", income_type_english),

```

```

      x = paste(income_type_english, "(euros)"),
      y = paste("Support for", party, "(%)"),
    ) +
    theme_minimal()
  }

```

Note that we have multiple different income metrics, and plotting all of them would be excessive. We can make a reasonable assumption that the average income after tax, i.e. the amount of money people actually get to keep, is the most important income measurement, so we will focus on this type of income for now. From this point onwards, any mention of "income" without specifying the type will, by default, refer to average income after tax. Let us plot income against support for the major parties:

```

> plot_income_against_support("Tulot_miinus_verot", "SDP")
> plot_income_against_support("Tulot_miinus_verot", "PS")
> plot_income_against_support("Tulot_miinus_verot", "KOK")
> plot_income_against_support("Tulot_miinus_verot", "KESK")
> plot_income_against_support("Tulot_miinus_verot", "VIHR")
> plot_income_against_support("Tulot_miinus_verot", "VAS")
> plot_income_against_support("Tulot_miinus_verot", "RKP")
> plot_income_against_support("Tulot_miinus_verot", "KD")
> plot_income_against_support("Tulot_miinus_verot", "LIIKE")

```

The above code generates the nine scatterplots in Fig 2. Note that we excluded the other parties because in most municipalities, their support amounts to not even 1% and thus it is difficult to notice any trend using such data. In the nine scatterplots, we noticed the following:

1. Support for the Finns Party mostly falls into the 0-40% range and has a slight negative correlation with income.
2. Support for the National Coalition Party mostly concentrates in the 0-30% range and has a quite strong positive correlation with income.
3. Support for the Centre Party varies vastly from 0% to nearly 60% and has a quite strong negative correlation with income.
4. Support for the Green League mostly ranges from 0% to 10% and seems to correlate positively with income.
5. Support for the Left Alliance mostly ranges from 0% to 20%, and there seems to be no correlation with income.
6. Support for the Swedish People's Party is near 0% in most municipalities. In the remaining few municipalities, however, their support can range from a few percentage points to over 75%. There also seems to be no correlation with income.
7. Support for the Christian Democrats mostly ranges from 0% to 10% and does not seem to correlate with income.
8. Support for Movement Now mostly ranges from 0% to 5% and seems to correlate positively with income.

Let us now look at the correlation between different income types and support for the above parties using a heatmap. We can draw the heatmap using the following code:

```

income_and_voting_columns <- c("Tulot", "Mediaanitulot", "Ansiotulot", "Pääomatulot",
  "Tulot_miinus_verot", "SDP", "PS", "KOK", "KESK", "VIHR", "VAS", "RKP", "KD", "LIIKE")
english_names <- c("Average Taxable Income", "Median Taxable Income", "Average Earned
  Income", "Average Investment Income", "Average Income after Tax", "SDP", "PS", "KOK",
  "KESK", "VIHR", "VAS", "RKP", "KD", "LIIKE")
cor_matrix <- cor(df[income_and_voting_columns])
cor_data <- as.data.frame(as.table(cor_matrix))
cor_data$Var1 <- factor(cor_data$Var1, levels = income_and_voting_columns, labels =
  english_names)

```

```

cor_data$Var2 <- factor(cor_data$Var2, levels = income_and_voting_columns, labels =
english_names)

ggplot(cor_data, aes(Var1, Var2, fill = Freq)) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(Freq, 2)), color = "black", size = 4) +
  scale_fill_gradient2(
    low = "blue",
    high = "red",
    mid = "white",
    midpoint = 0,
    limit = c(-1, 1),
    name = "Correlation"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    axis.title.x = element_blank(),
    axis.title.y = element_blank()
  )

```

The heatmap is shown in Fig. 3. In the heatmap, we noticed some of the previous correlations:

1. Municipalities with higher incomes appear to be more likely to support the National Coalition Party or the Green League. With the exception of average investment income, the Pearson correlation coefficient between the support for either party and any income metric is higher than 0.5.
2. Municipalities with lower incomes tend to support the Centre Party. With the exception of average investment income, the correlation coefficient between support for the party and income of any metric is lower than -0.6.

Let us now look at party support by electoral district by creating a stacked bar chart to show the composition of votes in each electoral district. Notice that we do not have the population for each municipality, so calculating the support for a certain party in each electoral district can be tricky. However, a reasonable assumption we can make is that the voting population in each municipality is roughly the same as the number of people who are working and paying tax in that municipality. Thus, we can calculate the average support for a certain party in a certain electoral district by weighing the district's constituent municipalities by their number of taxable income recipients. We do this using the following code:

```

parties <- c("SDP", "PS", "KOK", "KESK", "VIHR", "VAS", "RKP", "KD", "LIIKE")

weighted_districts <- df %>%
  group_by(Vaalipiiri) %>%
  summarize(across(all_of(parties), ~ weighted.mean(., Tulonsaajia)))

weighted_districts_long <- weighted_districts %>%
  pivot_longer(cols = all_of(parties), names_to = "Party", values_to = "Support")

ggplot(weighted_districts_long, aes(x = Vaalipiiri, y = Support, fill = Party)) +
  geom_bar(stat = "identity") +
  labs(
    x = "Electoral District",
    y = "Weighted Average Support (%)",
    fill = "Party"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme_minimal()

```

The resulting stacked bar chart is shown in Fig. 4. We noticed that:

1. Support for the National Coalition Party is strongest in the electoral districts of Helsinki and

Uusimaa, which are both located in Uusimaa, Finland's wealthiest and most urban region [2].

2. Support for the Centre Party is strongest in Lapland and Oulu, the two northernmost electoral districts of Finland. Most areas in these two districts are classified as rural by the Finnish Environment Institute [1]. Interestingly, support for the Finns Party also appears to be strongest in these two districts, and both party have roughly the same level of support in both districts.
3. The Social Democratic Party appears to have around a fifth of the votes in most electoral districts.
4. Support for the Swedish People's Party is insubstantial in most electoral districts. However, in Vaasa, their support amounts to around 20%.
5. No single party has a simple majority (50%) in any district. The most support a party has in one district is around 25%.

## 4 Formulating Hypotheses and Conducting Tests

Based on the above observations, we want to test if there is statistical evidence for the following hypotheses:

- Municipalities with higher incomes are more likely to support the National Coalition Party or the Green League.
- Municipalities with lower incomes are more likely to support the Finns party or the Centre Party.
- The National Coalition Party receives stronger support in urban areas, while the Centre Party and Finns Party receive stronger support in rural areas.
- Support for the Social Democratic Party appears to be around 20% in most electoral districts.
- Municipalities in Vaasa have substantially higher support for the Swedish People's Party than municipalities in other electoral districts.

### 4.1 Municipalities with higher incomes are more likely to support the National Coalition Party.

For this hypothesis, we can perform linear regression analysis. Assuming that there is a linear relationship between income and support for the National Coalition Party, then:

$Y = \beta_0 + \beta_1 X$ , where:

$Y$ : Support for the National Coalition Party

$X$ : Income

$H_0$ : There is no relationship between income and support for the National Coalition Party, i.e.,  $\beta_1 = 0$ .

$H_A$ : There is a positive relationship between income and support for the National Coalition Party, i.e.,  $\beta_1 > 0$ .

We can create the linear regression model as follows:

```
model <- lm(KOK ~ Tulot_miinus_verot, data = df)
summary(model)
```

The above code outputs:

Call:

```
lm(formula = KOK ~ Tulot_miinus_verot, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-14.7911	-4.1130	-0.3091	3.5252	24.9767

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2.206e+01	2.993e+00	-7.371	1.77e-12 ***
Tulot_miinus_verot	1.721e-03	1.393e-04	12.358	< 2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.211 on 291 degrees of freedom  
Multiple R-squared: 0.3442, Adjusted R-squared: 0.3419  
F-statistic: 152.7 on 1 and 291 DF, p-value: < 2.2e-16

The calculated slope is 1.721e-03, which is indeed positive. The coefficient of determination is 0.3442, meaning 34.42% of the variability in the support for the National Coalition Party can be explained by income. We can fit the linear regression line using the following code:

```
ggplot(data = df, mapping = aes(x = Tulot_miinus_verot, y = KOK)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(
    x = "Average Income after Tax (euros)",
    y = "Support for KOK (%)",
  ) +
  theme_minimal()
```

The fitted linear regression line is shown in Fig. 5.

As can be seen in the figure, the line indeed has a clear positive slope.

The two-sided p-value is less than 2e-16, which is extremely small, so the one-sided p-value would be lower than any commonly used threshold, and as such we reject  $H_0$  and accept  $H_A$ . In other words, there is significant statistical evidence to suggest that municipalities with higher income are more likely to support the National Coalition Party.

## 4.2 Municipalities with higher incomes are more likely to support the Green League.

Again, we shall perform linear regression analysis.

$H_0$ : There is no relationship between income and support for the Green League, i.e.,  $\beta_1 = 0$ .

$H_A$ : There is a positive relationship between income and support for Green League, i.e.,  $\beta_1 > 0$ .

The code is almost the same as before:

```
model <- lm(VIHR ~ Tulot_miinus_verot, data = df)
summary(model)
```

Applying the linear regression model gives the following results:

Call:

```
lm(formula = VIHR ~ Tulot_miinus_verot, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.2786	-1.0822	-0.3528	0.6284	9.2370

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-7.186e+00	8.838e-01	-8.131	1.24e-14 ***
Tulot_miinus_verot	4.815e-04	4.112e-05	11.710	< 2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.834 on 291 degrees of freedom



Multiple R-squared: 0.3203, Adjusted R-squared: 0.3179  
F-statistic: 137.1 on 1 and 291 DF, p-value: < 2.2e-16

The calculated slope is 4.815e-04, which is positive. The coefficient of determination is 0.3203, meaning that roughly 32% of the variability in the support for the Green League can be explained by income. The fitted regression line is shown in Fig. 6.

Again, since the two-sided p-value is less than 2e-16, which is extremely small, the one-sided p-value would also be lower than any commonly used threshold, so we reject  $H_0$  and accept  $H_A$ . In other words, there is significant statistical evidence to suggest that there is a positive relationship between a municipality's average income after taxes and their support for the Green League.

### 4.3 Municipalities with lower incomes are more likely to support the Finns Party.

$H_0$ : There is no relationship between income and support for the Finns Party, i.e.,  $\beta_1 = 0$ .

$H_A$ : There is a negative relationship between income and support for the Finns Party, i.e.,  $\beta_1 < 0$ .

As before, we perform linear regression analysis:

```
> model <- lm(PS ~ Tulot_miinus_verot, data = df)
> summary(model)
Call:
lm(formula = PS ~ Tulot_miinus_verot, data = df)

Residuals:
    Min       1Q   Median       3Q      Max
-23.5736  -3.1454   0.0858   3.6998  27.6090

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.4438263   3.2467606   12.149  < 2e-16 ***
Tulot_miinus_verot -0.0006924  0.0001511  -4.583  6.8e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 6.737 on 291 degrees of freedom  
Multiple R-squared: 0.06732, Adjusted R-squared: 0.06412  
F-statistic: 21.01 on 1 and 291 DF, p-value: 6.804e-06

The calculated slope is -0.0006924, which is indeed negative. The coefficient of determination is 0.06732, indicating that 6.7% of the variability in support for the Finns Party can be explained by income. Since the two-sided p-value, 6.8e-06, is already incredibly small, we reject the null hypothesis and conclude that there is a negative relationship between income and support for the Finns Party. However, since both the slope of the linear regression line and the coefficient of determination are small, this negative relationship is not a strong one. In other words, the effect of income on the support for the Finns Party is negative but limited.

### 4.4 Municipalities with lower incomes are more likely to support the Centre Party.

$H_0$ : There is no relationship between income and support for the Centre Party, i.e.,  $\beta_1 = 0$ .

$H_A$ : There is a negative relationship between income and support for the Centre Party, i.e.,  $\beta_1 < 0$ .

Performing linear regression:

```
> model <- lm(KESK ~ Tulot_miinus_verot, data = df)
> summary(model)
Call:
lm(formula = KESK ~ Tulot_miinus_verot, data = df)

Residuals:
```

Min	1Q	Median	3Q	Max
-25.681	-6.390	-0.056	6.163	50.077

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	89.5518042	4.9520159	18.08	<2e-16 ***
Tulot_miinus_verot	-0.0031380	0.0002304	-13.62	<2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.28 on 291 degrees of freedom  
Multiple R-squared: 0.3893, Adjusted R-squared: 0.3872  
F-statistic: 185.5 on 1 and 291 DF, p-value: < 2.2e-16

The coefficient of determination is 0.3893, indicating that 38.93% of the variability in the support for the Centre Party can be explained by average income after taxes. The two-sided p-value is extremely small (less than 2e-16), so the one-sided p-value is even smaller, and we again reject the null hypothesis. Indeed, the linear regression line has a clear downwards slope, as shown in Fig. 7.

#### 4.5 The National Coalition Party receives stronger support from urban areas.

Before we can start testing this hypothesis, we need to classify municipalities as either urban or rural. We shall use the list at [https://en.wikipedia.org/wiki/List\\_of\\_urban\\_areas\\_in\\_Finland\\_by\\_population](https://en.wikipedia.org/wiki/List_of_urban_areas_in_Finland_by_population) as reference; any municipality included in this list will be considered urban and every other municipality will be considered rural.

Classifying municipalities:

```
urban_areas <- c(
  "Helsinki", "Tampere", "Turku", "Oulu", "Jyväskylä", "Lahti", "Kuopio", "Pori",
  "Joensuu", "Vaasa", "Lappeenranta", "Rovaniemi", "Seinäjoki", "Hämeenlinna",
  "Porvoo", "Kotka", "Kouvola", "Hyvinkää", "Mikkeli", "Kokkola", "Rauma", "Lohja",
  "Kajaani", "Salo", "Riihimäki", "Imatra", "Kemi", "Forssa", "Jakobstad",
  "Savonlinna", "Kirkkonummi", "Raahe", "Varkaus", "Valkeakoski", "Tornio",
  "Hamina", "Iisalmi", "Mariehamn", "Nummela", "Heinola", "Ilmajoki", "Kurikka",
  "Pieksämäki", "Ylivieska", "Jämsä", "Nastola", "Mäntsälä", "Siilinjärvi", "Lapua",
  "Uusikaupunki", "Vammala", "Söderkulla", "Pargas", "Orimattila", "Loimaa", "Ekenäs",
  "Kauhajoki", "Äänekoski", "Paimio", "Toijala", "Kuusamo", "Laukaa", "Karis",
  "Kankaanpää", "Nurmijärvi", "Turenki", "Mänttä", "Karkkila", "Hanko",
  "Rajamäki", "Muurame", "Muhos", "Loviisa", "Liekka", "Joutseno", "Kyröskoski",
  "Parola", "Lauttakylä", "Laihia", "Kalajoki", "Iin Hamina", "Jokela", "Eura",
  "Orivesi", "Veikkola", "Kyläsaari", "Pihlava", "Vuokatti", "Keuruu", "Valkeala",
  "Myllykoski", "Kiiminki", "Laitila", "Toivala", "Vuorela", "Kauhava", "Vuoress",
  "Nivala", "Oulainen", "Kuhmo", "Liminka", "Viiala", "Suonenjoki"
)
```

```
df$Type <- ifelse(df$Alue %in% urban_areas, "urban", "rural")
```

Let us denote  $\mu_{urban}$  as the mean level of support (for the National Coalition Party, in this case) across all urban municipalities and  $\mu_{rural}$  as the mean level of support for the party across all rural municipalities.

$H_0$ : There is no difference in the support for the National Coalition Party between urban and rural municipalities, i.e.  $\mu_{urban} = \mu_{rural}$ .

$H_A$ : Support for the National Coalition Party is greater in urban municipalities, i.e.,  $\mu_{urban} > \mu_{rural}$ .

We can use a 2-sample t-test in this case:

```
> urban_support <- df$KOK[df$Type == "urban"]
> rural_support <- df$KOK[df$Type == "rural"]
```

```
> result <- t.test(urban_support, rural_support, alternative = "greater", conf.level
= 0.99)
> print(result)
```

Welch Two Sample t-test

```
data: urban_support and rural_support
t = 3.9402, df = 168.36, p-value = 5.954e-05
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 1.384947      Inf
sample estimates:
mean of x mean of y
 17.22877  13.80000
```

The p-value, 5.954e-05, is much smaller than any common significance value. The calculated average support for the National Coalition Party across urban municipalities, 17.23%, is also reasonably higher than the average support for the party across rural municipalities, 13.8%. The 99% confidence interval is entirely positive and does not contain zero either. Thus, we reject the null hypothesis and accept the alternative hypothesis.

#### 4.6 The Centre Party receives stronger support from rural areas.

$H_0$ : There is no difference in the support for the Centre Party between urban and rural municipalities, i.e.  $\mu_{urban} = \mu_{rural}$ .

$H_A$ : Support for the Centre Party is greater in rural municipalities, i.e.,  $\mu_{urban} < \mu_{rural}$ .

Again, we shall use a 2-sample t-test:

```
> urban_support <- df$KESK[df$Type == "urban"]
> rural_support <- df$KESK[df$Type == "rural"]
> result <- t.test(urban_support, rural_support, alternative = "less", conf.level =
0.99)
> print(result)
```

Welch Two Sample t-test

```
data: urban_support and rural_support
t = -4.824, df = 147.37, p-value = 1.738e-06
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
 -Inf -3.868026
sample estimates:
mean of x mean of y
 16.93836  24.48636
```

The average support for the Centre Party across urban municipalities is about 16.94%, but in rural municipalities the figure is much higher at 24.49%. Since the 99% confidence interval does not include zero and the p-value, 1.738e-06, is much less than our significance level of 0.01, we reject the null hypothesis.

#### 4.7 The Finns Party receives stronger support from rural areas.

$H_0$ :  $\mu_{urban} = \mu_{rural}$

$H_A$ :  $\mu_{urban} < \mu_{rural}$

Again, we use a 2-sample t-test:

```
> urban_support <- df$PS[df$Type == "urban"]
> rural_support <- df$PS[df$Type == "rural"]
> result <- t.test(urban_support, rural_support, alternative = "less", conf.level =
0.99)
```

```
> print(result)
```

Welch Two Sample t-test

```
data: urban_support and rural_support
t = -0.54051, df = 157.19, p-value = 0.2948
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
 -Inf 1.504662
sample estimates:
mean of x mean of y
 24.33562  24.78500
```

The p-value, 0.2948, is not smaller than our significance level of 0.01. The average support for the Finns Party across urban municipalities is 24.34%, which is virtually the same as the average support for the party across rural municipalities. Thus, we fail to reject  $H_0$ .

#### 4.8 Support for the Social Democratic Party is uniform across electoral districts at 20%.

Let us denote  $\mu$  as the average support for the SDP across electoral districts.

$H_0: \mu = 20\%$   
 $H_A: \mu \neq 20\%$

We shall use a one-sample t-test:

```
> weighted <- df %>%
+   group_by(Vaalipiiri) %>%
+   summarise(weighted_sdp_support = sum(SDP * Tulonsaajia) / sum(Tulonsaajia))
> result <- t.test(weighted$weighted_sdp_support, mu = 20, conf.level = 0.99)
> print(result)
```

One Sample t-test

```
data: weighted$weighted_sdp_support
t = 0.16673, df = 11, p-value = 0.8706
alternative hypothesis: true mean is not equal to 20
99 percent confidence interval:
 16.28294 24.13878
sample estimates:
mean of x
 20.21086
```

The p-value is 0.8706, which is not less than our significance level of 0.01. Thus, we should fail to reject the null hypothesis. In other words, we cannot conclude that SDP has uniform support at 20% at every electoral district.

#### 4.9 Municipalities in Vaasa have substantially higher support for the Swedish People's Party than municipalities in other electoral districts.

Let us denote  $\mu_{Vaasa}$  as the average unweighted support for the Swedish People's Party across municipalities in Vaasa and  $\mu_{Other}$  as the average unweighted support for the party across municipalities in other electoral districts.

$H_0: \mu_{Vaasa} = \mu_{Other}$   
 $H_A: \mu_{Vaasa} > \mu_{Other}$

We shall use a 2-sample t-test:

```
> vaasa <- df$RKP[df$Vaalipiiri == "Vaasa"]
> other <- df$RKP[df$Vaalipiiri != "Vaasa"]
> result <- t.test(vaasa, other, alternative = "greater", conf.level = 0.99)
> print(result)
```

Welch Two Sample t-test

```
data: vaasa and other
t = 3.6727, df = 39.87, p-value = 0.0003523
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 5.939094      Inf
sample estimates:
mean of x mean of y
19.580000  2.117391
```

The p-value is 0.0003523, much smaller than our significance level of 0.01. Thus, we reject the null hypothesis and accept the alternative hypothesis that the Swedish People's Party receives much more support in Vaasa's municipalities than municipalities in other electoral districts.

## 5 Summary

In this report, we analysed Finnish citizens' voting behaviour by taking into account their income by various metrics across municipalities countrywide in addition to their geographic locations. We found statistically significant evidence, at 99% confidence, that:

1. Municipalities with higher income are more likely to support the National Coalition Party or the Green League.
2. There is a negative but very limited relationship between a municipality's income and their support for the Finns Party.
3. Municipalities with lower income are also more likely to support the Centre Party.
4. The National Coalition Party tends to receive higher support in urban municipalities.
5. The Centre Party, on the other hand, receives higher support in rural municipalities.
6. The Swedish People's Party receives significantly more support in the municipalities of Vaasa than in the municipalities of other electoral districts.

## References

- [1] Finnish Environment Institute. *Updated urban-rural classification: Finland's degree of urbanisation currently at over 72 per cent*. Accessed: 2024-09-21. 2020. URL: [https://www.syke.fi/en-US/Current/Updated\\_urbanrural\\_classification\\_Finlan\(57443\)](https://www.syke.fi/en-US/Current/Updated_urbanrural_classification_Finlan(57443)).
- [2] Statista. *Gross Domestic Product (GDP) per capita in Finland in 2021, by region*. Accessed: 2024-09-21. 2021. URL: <https://www.statista.com/statistics/1150699/finland-gross-domestic-product-gdp-per-capita-by-region/>.

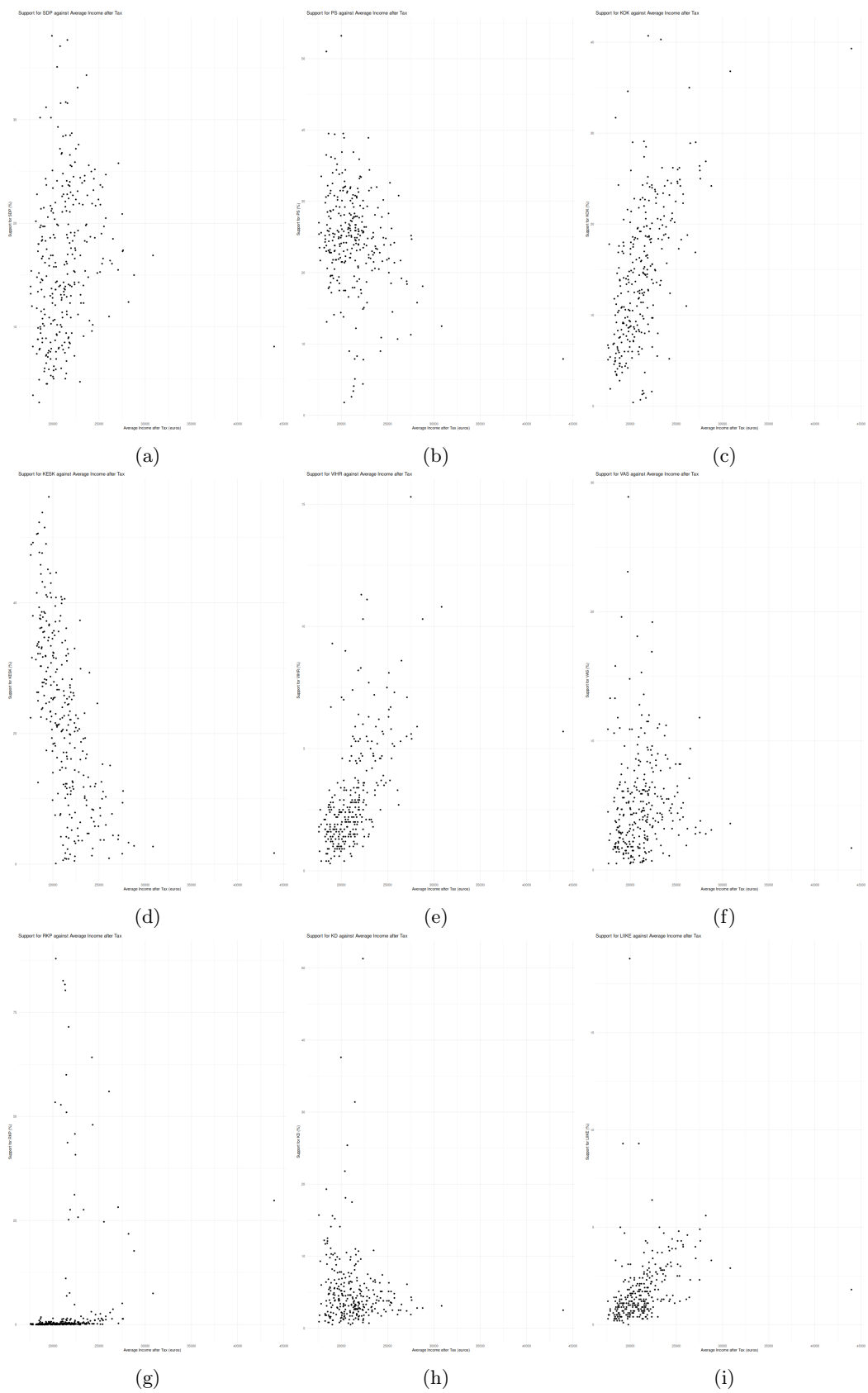


Figure 2: Support for each major party against Average Income after Tax

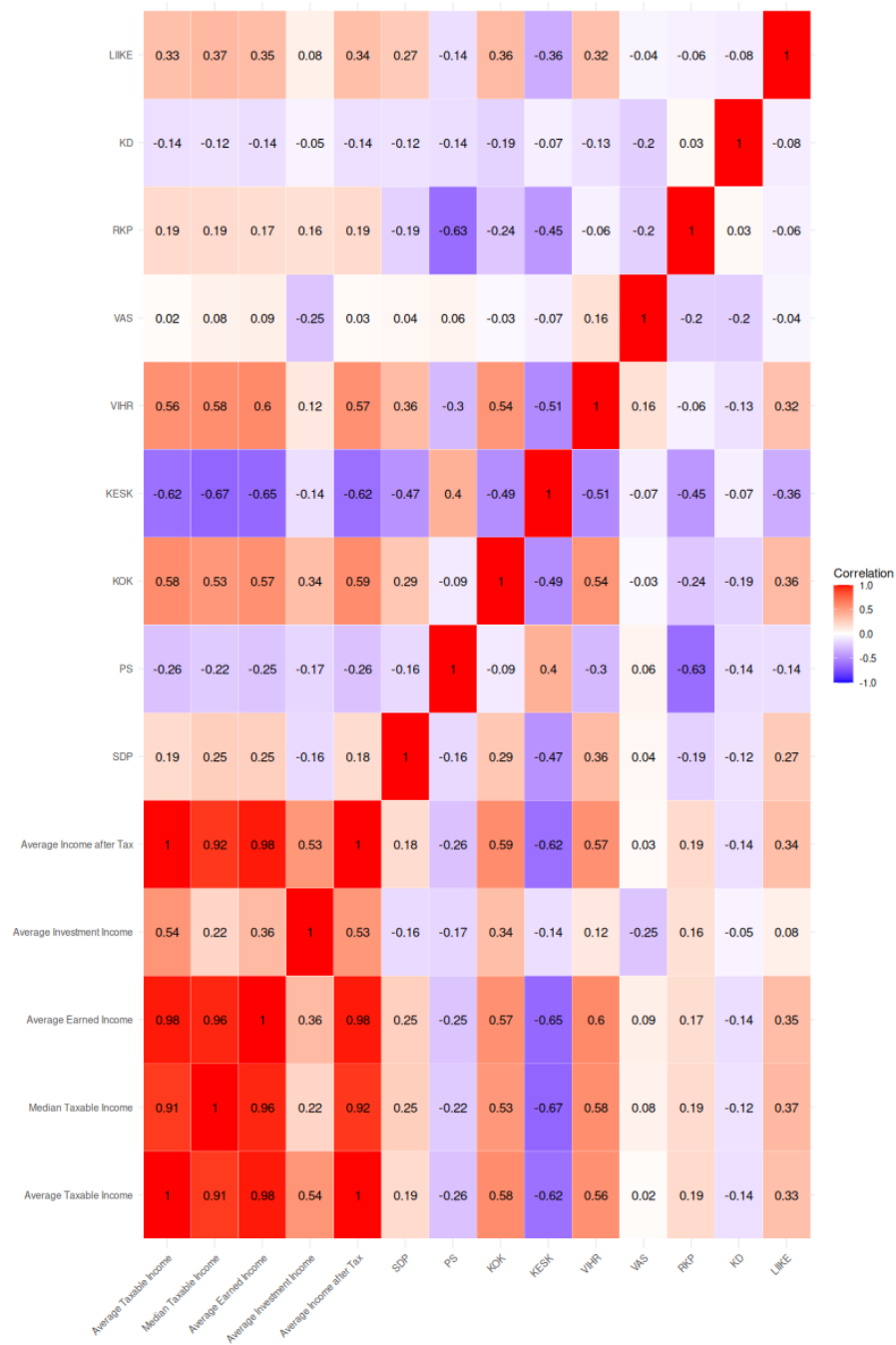


Figure 3: Correlation Heatmap: Income Metrics vs Party Support

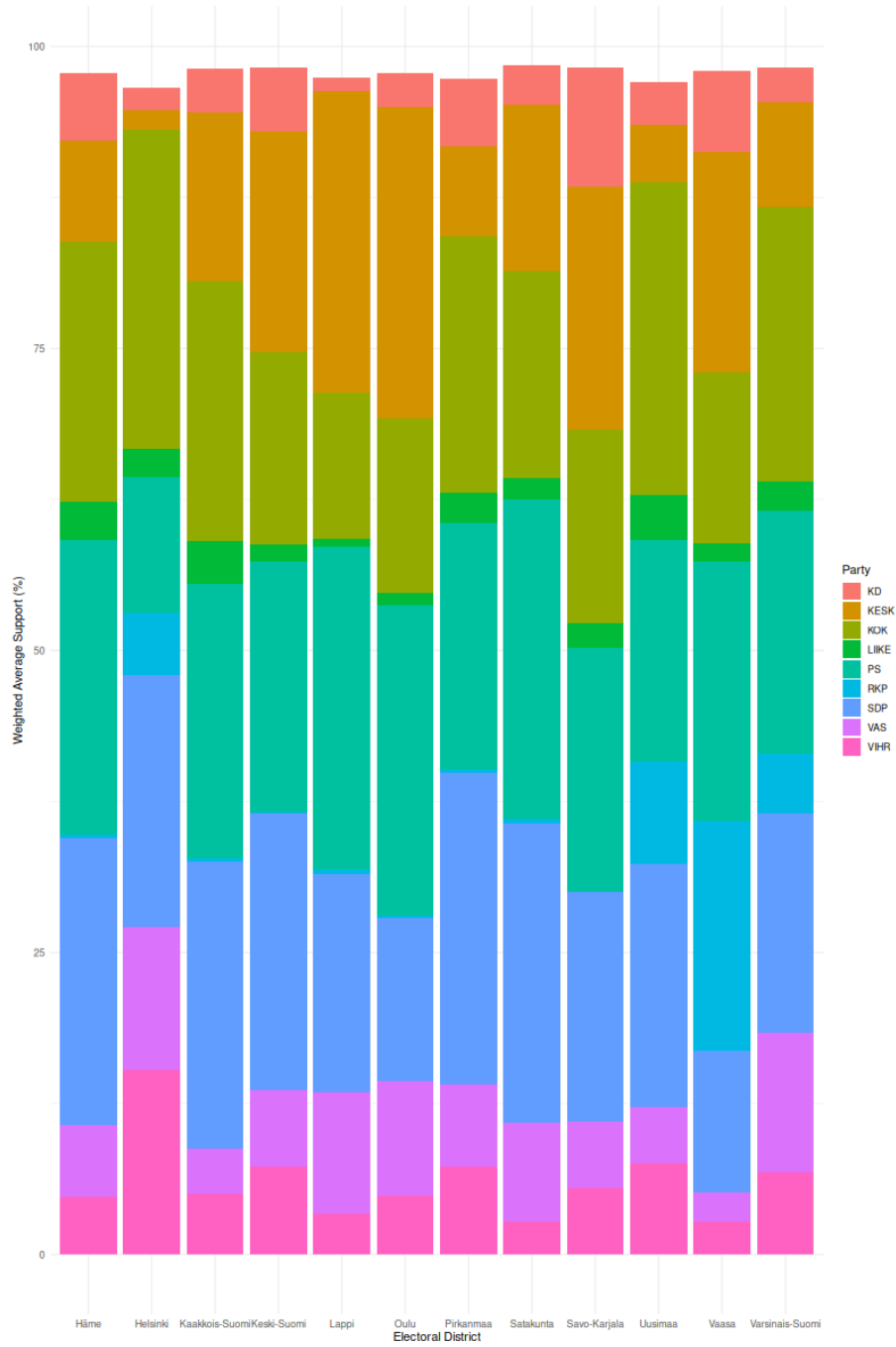


Figure 4: Population-Weighted Average Voting Support by Electoral District



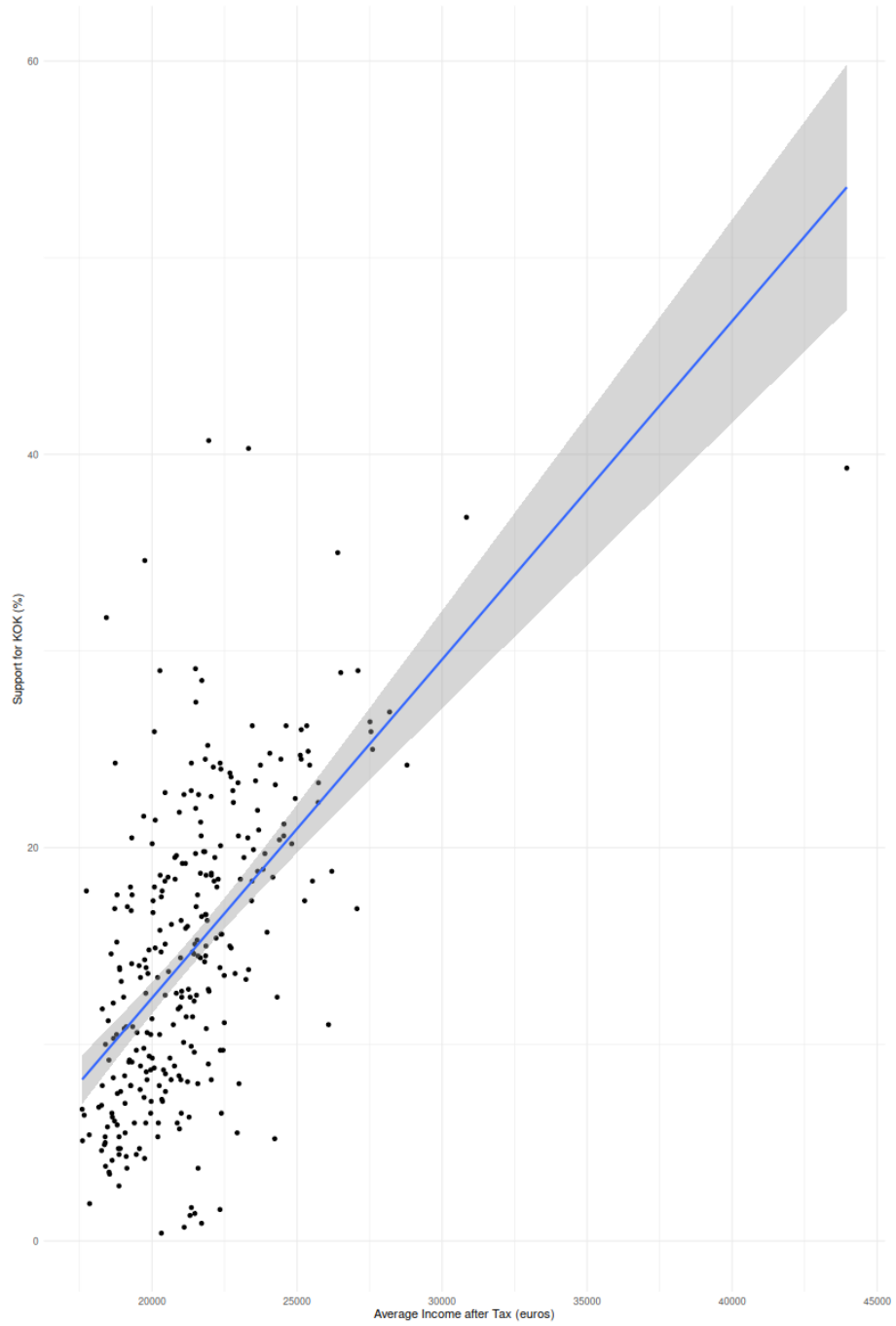


Figure 5: Support for KOK against Average Income after Tax

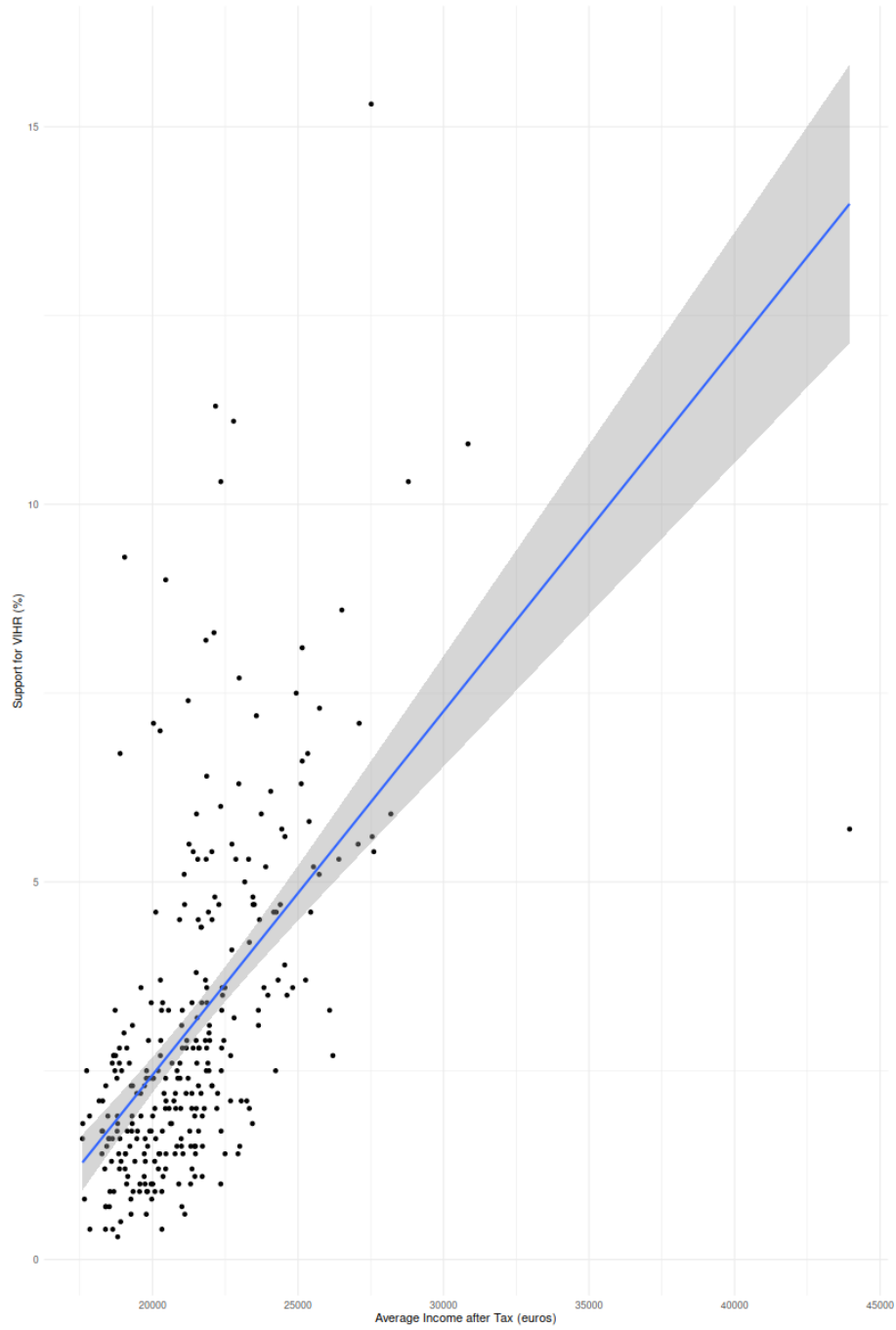


Figure 6: Support for VIHR against Average Income after Tax

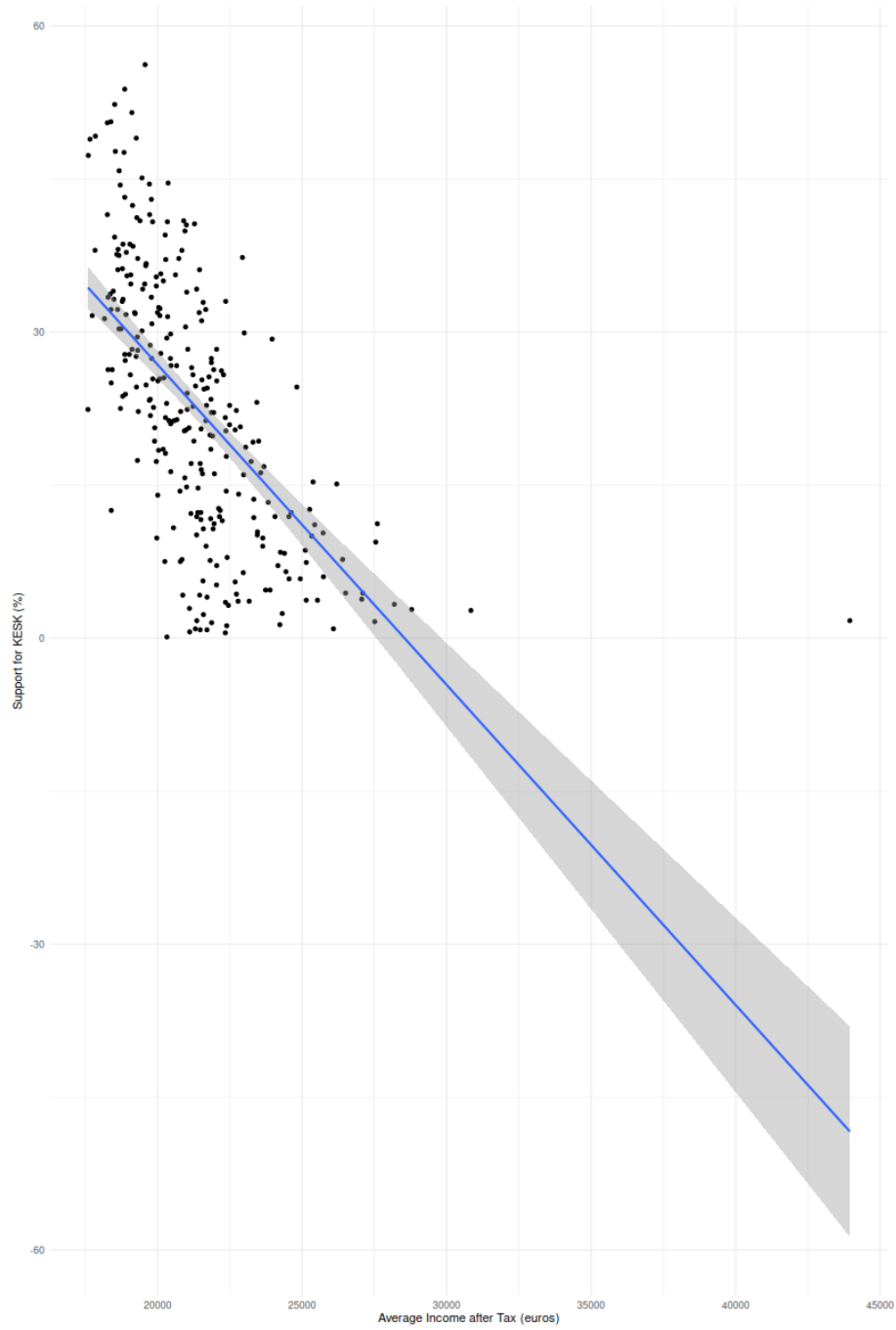


Figure 7: Support for KESK against Average Income after Tax