RELATIONSHIP OF INCOME LEVEL AND METROPOLITAN AREA WITH VOTING BEHAVIORS IN FINLAND

I. Introduction

This report is interested in finding how the income levels and metropolitan areas are related to voting behaviors in Finland. The voting behavior in question refers to party vote choices. This report's objective is to find out if there is any trend among certain municipalities with certain levels of income towards the municipalities' choices of politic parties in Finland.

II. Data

Prior to the analysis, here are some disclaimers:

Issues & Limitations:

1. The number of municipalities slightly differs in the two datasets.
2. The income data is from 2017, but the election data is from 2019.

Important to note:

1. Each municipality is independent from the others.
2. The variables in the income statistics are condensed indicators (consisting of averages and medians) of municipal individuals.

2.1 Data Sources

There are in total 2 (two) datasets that are used in this analysis.

The data file *ek2019*.csv contains the result of 2019 parliamentary elections in Finland, categorized by municipality, and a constituency to which the voting area belongs to is also included. The data includes the percentage of voting choice of political party in each municipality. The parties in the dataset include: Suomen Keskusta (KESK), Perussuomalaiset (PS), Kansallinen Kokoomus (KOK), Suomen Sosiaalidemokraattinen puolue (SDP), Vihreä liitto (VIHR), Vasemmistoliitto (VAS), Suomen ruotsalainen Kansanpuolue (RKP), Suomen Kristillisdemokraatit (KD), Suomen Kommunistinen Puolue (SKP), Kommunistinen Työväenpuolue - Rauhan ja Sosialismin puolesta (KTP), Liberaalipuolue - Vapaus valita (LIB), Piraattipuolue (PIR), Eläinoikeuspuolue (EOP), Kansalaispuolue (KP), Feministinen puolue (FP), Itsenäisyyspuolue (IP), Sininen tulevaisuus (SIN), Suomen Kansa Ensin (SKE), Seitsemän tähden liike (STL), Valitsijayhdistykset (Muut).

The data file *tulot2017*.csv contains information about income in Finland, which includes numeric representations of: taxable income recipients in the area (*Tulonsaajia*), average taxable income of the region's income recipients in euros (*Tulot*), median taxable income of the income recipients in the region in euros (*Mediaanitulot*), average earned income of the area's income recipient in euros (*Ansiotulot*), average investment (capital) income of the area's income recipients in euros (*Pääomatulot*), average state tax paid by the income recipients of the region in euros (*Valtionvero*), average municipal tax paid by income recipients of the area in euros (*Kunnallisvero*), total average tax for income recipients in the area (*Verot*), average income after tax of the income recipients in the area (*Tulot miinus verot*). These are categorized per municipality as well.

The income dataset has all the municipalities in Finland, however, the election dataset does not.

2.2 Data Cleaning

The objective of this analysis is to find out the relationship of income levels and municipalities with voting behaviors. Due to the nature of statistical analysis, it is a wise decision to leave out the missing municipalities, as there can be no comparison to the income dataset if the data of some municipalities is missing in the election dataset. In result, this analysis will not be able to cover the whole Finland.

After the cleaning process, municipalities that exist in the income datasets but not in the election dataset are found. Those municipalities are: Brändö, Eckerö, Finström, Föglö, Geta, Hammarland, Jomala, Kumlinge, Kökar, Lemland, Lumparland, Maarianhamina, Saltvik, Sottunga, Sund, Vårdö. These are all municipalities of Åland (Ahvenanmaa). These municipalities will be removed prior to the analysis.

Cleaning steps:

1. Unique values from the "*Alue*" (Municipality) column of each dataset were extracted.
2. The unique values from each dataset were combined, and the above-mentioned results were found.

There is also a "*KOKO MAA*" (whole country) row in "*Alue"* (Municipality) column of income dataset which includes the data for the whole country. However, it will not be used because there is no comparison of the whole country in the election dataset.

The final data processing done is to merge the two datasets, discarding the above-mentioned differences in "*Alue*" (Municipality) column. The merged dataset now consists of 31 total columns and 295 total entries.

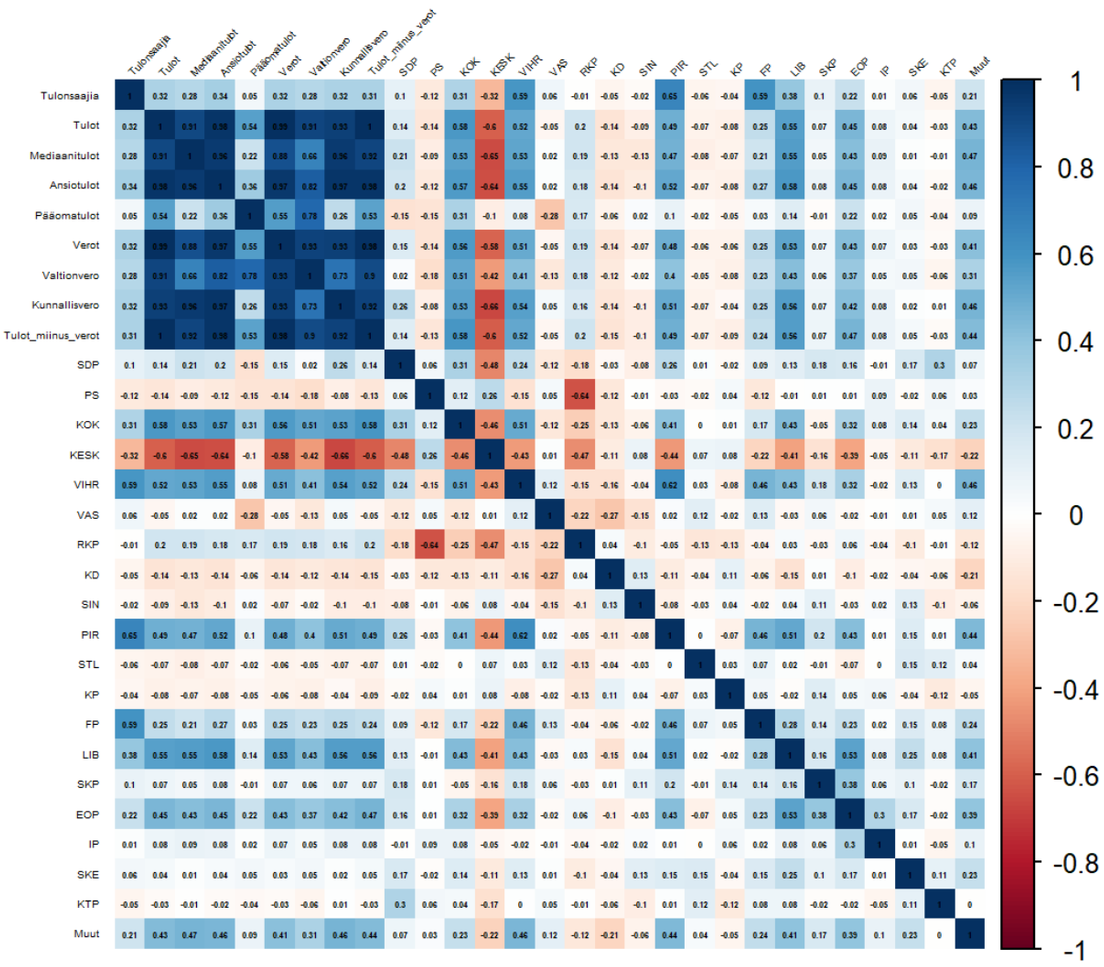
III. Summary, Visualization, and Analysis

3.1 Correlation Test of Income Levels and Voting Decisions

This section is going to identify if there is a trend in which a certain income level prefers a certain political party. As a first step, correlation matrix between all the variables of the merged dataset is made to quickly extract which variables have significant correlations. It is the first hint to see how closely related the variables are. For reference, Mukaka (2012)[1]'s rule of determining correlation coefficient is utilized:

Absolute Size of Correlation	Interpretation
.90 – 1.00	Very high correlation
.70 – .90	High correlation
.50 – .70	Moderate correlation
.30 – .50	Low correlation
.00 – .30	Negligible correlation

Here is the visualization of the correlation matrix:

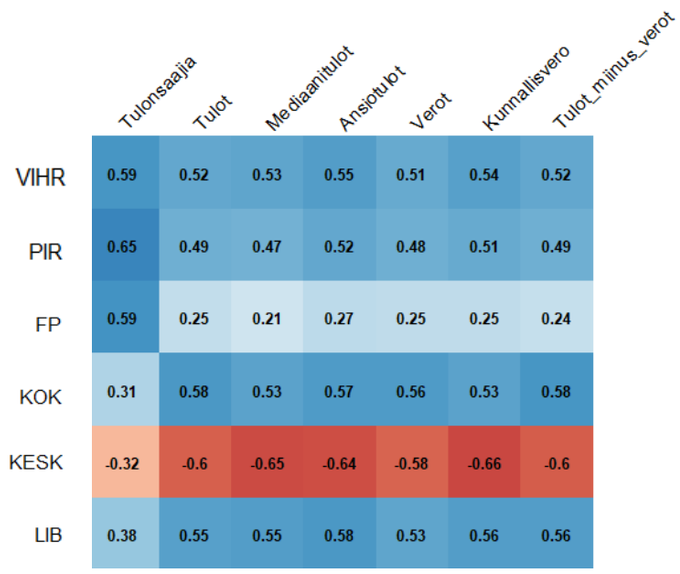


The visualization of the correlation matrix is colored in shades of blue and red. The redder it is, the lower the correlation coefficient. The bluer it is, the higher the correlation coefficient. To aid, the correlation coefficient is also explicitly written. Right away, with the help of color indicators and text, we see which variables are closely related to the other. However, it is important to note that coefficients of income and

[1] Mukaka, MM. "A guide to appropriate use of Correlation coefficient in medical research." Malawi Medical Journal 24, no. 3 (September 2012): 69-71

party turnouts only serve as an indication of correlation, not causation. Therefore, there needs to be further examination.

In the correlation matrix, there seems to be a highlight – the same set of parties keeps being associated to certain income variables. Here is the summarized data from the matrix:

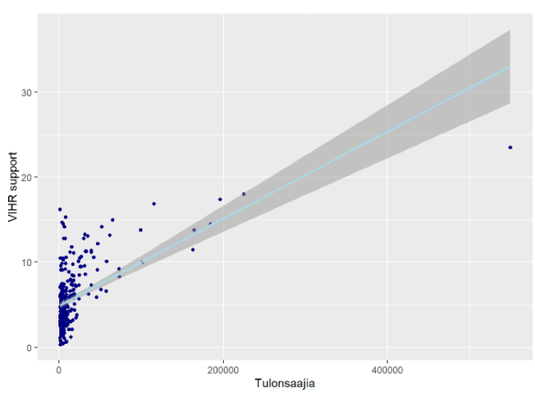


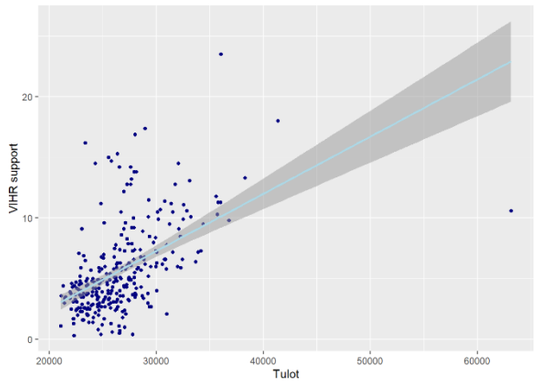
The above parties have few significant correlations with the income variables. To further statistically verify that they do have significant relationship, a simple linear regression will be used. A simple linear regression is used to estimate the relationship between two quantitative variables, in this case, party support and income variables. The data exploration will focus per party and take the perspective of each party. This way, we will be able to focus on the behavior of certain municipalities with certain income levels towards choosing a political party. In this report, only significant relationships will be shown. For a quick reference, in a simple linear regression, this report uses these indicators to determine the significance of the relationship:

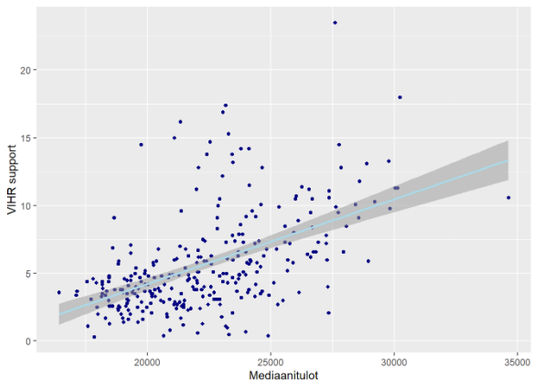
- R^2 = a measure of how well a linear regression model fits a dataset. Ranging from 0 to 1, a value of 0 indicates the response variable cannot be explained by the predictor, and a value of 1 indicates the response variable can be perfectly explained by the predictor. In explaining the relationship between predictor and response variable (which is what we are doing), R^2 value of 0.2 is already considered high especially for real-world data.
- P-value = a number that describes how likely it is to find a particular set of observations if the null hypothesis (there is no relationship between the variables) was true. The lower the p-value, the more likely it is that the variables are statistically significant. The most common threshold of the p-value to be called small enough is 0.05.

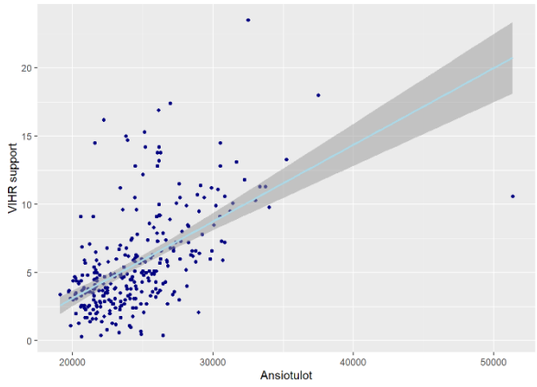
Before the analysis per party, it is also important to note how closely related the income variables are. It is apparrent that average taxable income, taxable median income, average earned income, average state tax, average municipal tax, total average taxes all are related in a positive direction. With such high correlation coefficients, it can safely be said that all the mentioned variables will increase when the other increases and decrease when the other decreases. In other scenarios where all the variables are very obviously unrelated to each other, this kind of assumption is unlikely to happen (because again, correlation does not mean causation). However, this dataset is already compiled so that all the variables are in the same context and therefore related to each other. In this case, if all the above-mentioned variables are high for a region, it can be concluded that on average, the said municipality have high income levels.

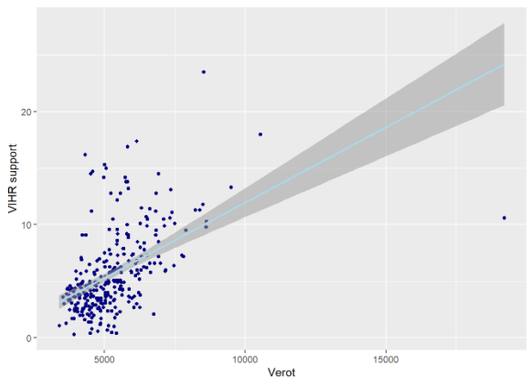
3.2 VIHR (Vihreä liitto)

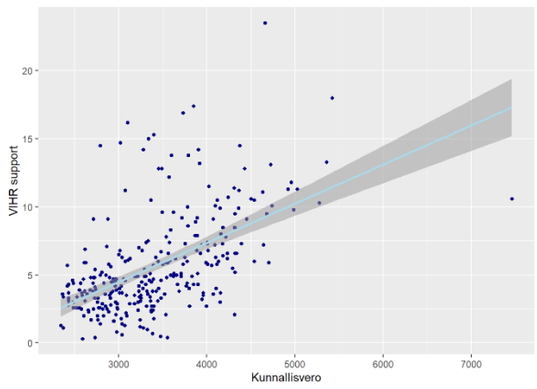


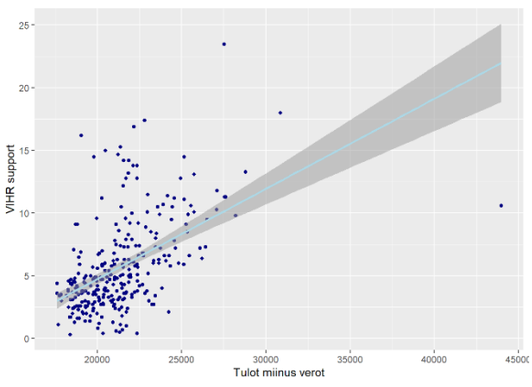








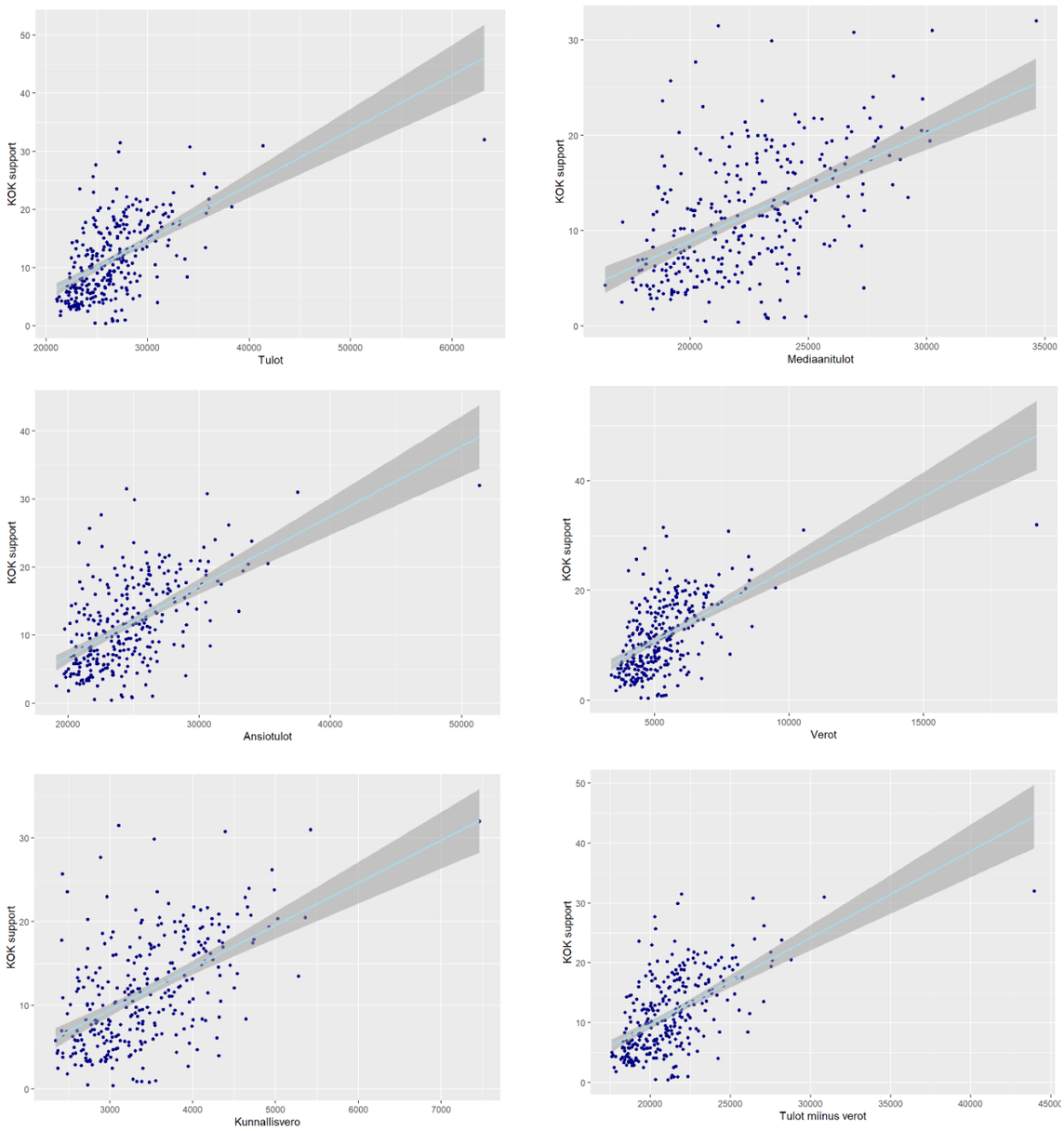




The above scatterplots (in order: *Tulonsaajia* (The number of taxable income recipients in the area, $R^2 = 0.347$, $p < 0.0001$), *Tulot* (average taxable income of the income recipients in the region in euros, $R^2 = 0.2664$, $p < 0.0001$), *Mediaanitulot* (taxable median income of the income recipients in the region in euros, $R^2 = 0.2762$, $p < 0.0001$), *Ansiotulot* (average earned income of the area's income recipients in euros, $R^2 = 0.3066$, $p < 0.0001$), *Verot* (total average taxes for income recipients in the area, $R^2 = 0.2569$, $p < 0.0001$), *Kunnallisvero* (average municipal tax paid by income recipients of the area in euros, $R^2 = 0.2937$, $p < 0.0001$), *Tulot miinus verot* (The average income after the taxes of the income recipients in the area, $R^2 = 0.2678$, $p < 0.0001$)) suggests that when each of those variables increase, VIHR support also tends to increase. There are a few outliers in every scatterplot, however, they still follow the positive trend.

Based on these findings, there are a few interpretations that could be made. Where the VIHR support is high, the number of taxable income recipients is also high. This could be indicating that those municipalities which have more individuals that are eligible to receive taxable income (not elders, children, or immigrants without work permit) tend to vote for VIHR party. When average taxable income, taxable median income, average municipal tax, total average taxes, and average income after taxes are high, VIHR support is also high. In this context, a conclusion that surfaces is that the municipalities where incomes are higher, where people are "richer" tend to vote for VIHR party.

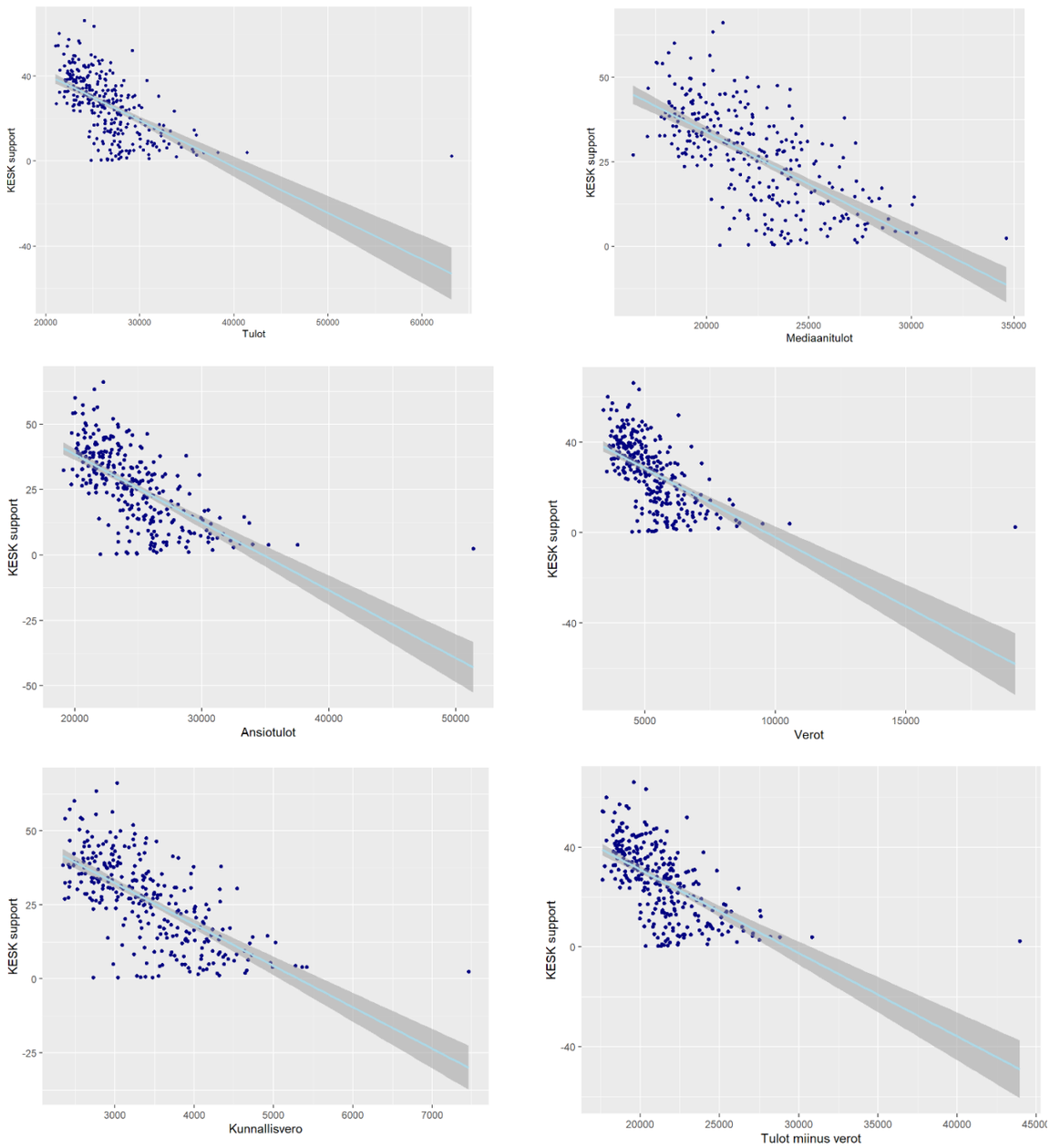
3.3 KOK (Kansallinen Kokoomus)



The above scatterplots (in order: *Tulot* (average taxable income of the income recipients in the region in euros, $R^2 = 0.3338$, $p < 0.0001$), *Mediaanitulot* (taxable median income of the income recipients in the region in euros, $R^2 = 0.2816$, $p < 0.0001$), *Ansiotulot* (average earned income of the area's income recipients in euros, $R^2 = 0.3197$, $p < 0.0001$), *Verot* (total average taxes for income recipients in the area, $R^2 = 0.314$, $p < 0.0001$), *Kunnallisvero* (average municipal tax paid by income recipients of the area in euros, $R^2 = 0.2846$, $p < 0.0001$), *Tulot miinus verot* (The average income after the taxes of the income recipients in the area, $R^2 = 0.3399$, $p < 0.0001$)) suggests that when each of those variables increase, KOK support also tends to increase. There are a few outliers in every scatterplot, however, they still follow the positive trend.

Just like the previous VIHR party, the trends of income variables compared to KOK party are similar. For instance, when the average taxable income and taxable median income, average earned income, average municipal tax, average tax, as well as the average income after taxation of the municipality are high, KOK support is also high. This indicates high-earning municipalities tend to vote for KOK. Like VIHR party, it can be concluded that KOK party is mostly preferred by those with higher income levels – "richer" populations.

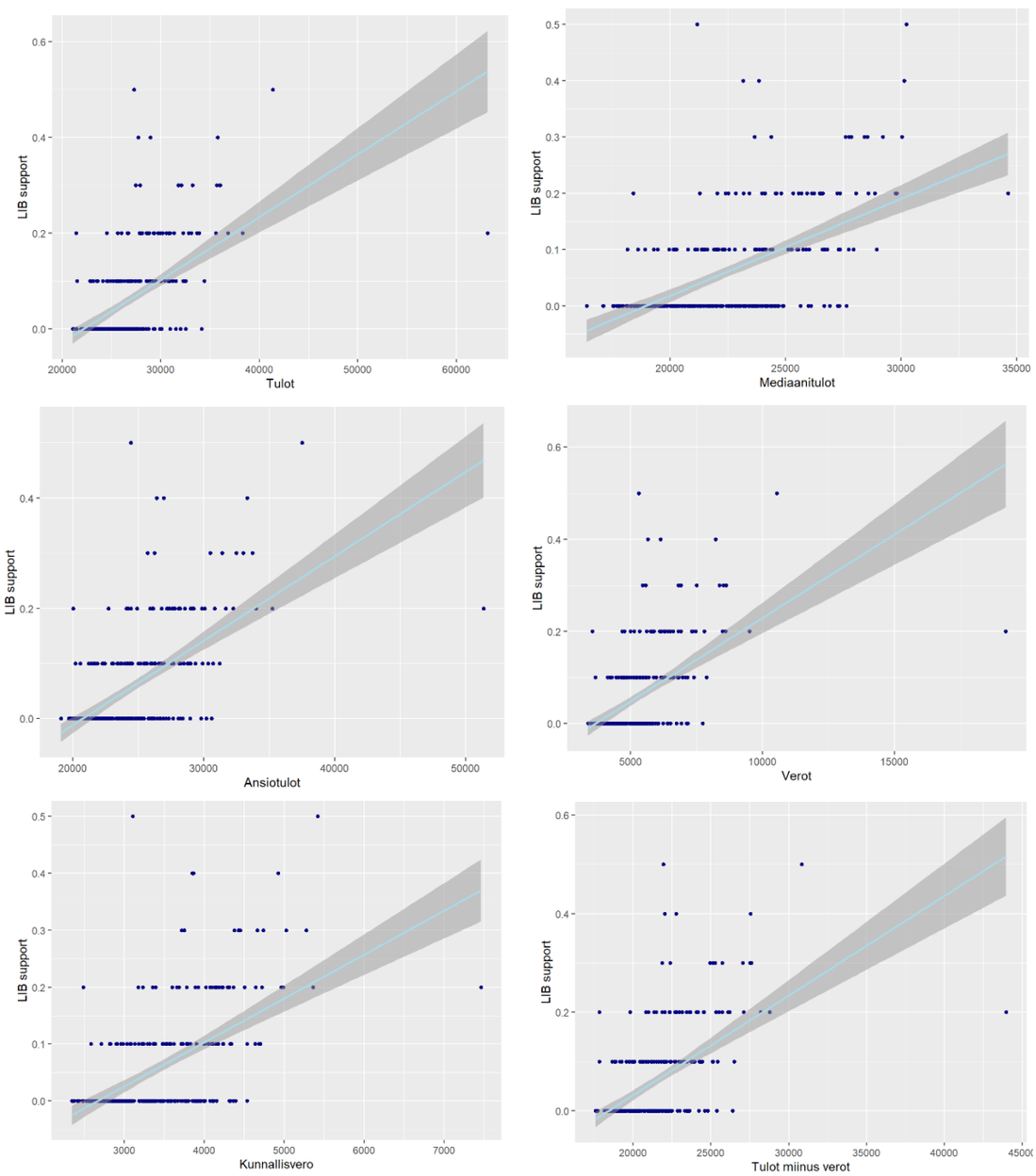
3.4 KESK (Suomen Keskusta)



The above scatterplots (in order: *Tulot* (average taxable income of the income recipients in the region in euros, $R^2 = 0.3583$, $p < 0.0001$), *Mediaanitulot* (taxable median income of the income recipients in the region in euros, $R^2 = 0.4249$, $p < 0.0001$), *Ansiotulot* (average earned income of the area's income recipients in euros, $R^2 = 0.4119$, $p < 0.0001$), *Verot* (total average taxes for income recipients in the area, $R^2 = 0.3392$, $p < 0.0001$), *Kunnallisvero* (average municipal tax paid by income recipients of the area in euros, $R^2 = 0.438$, $p < 0.0001$), *Tulot miinus verot* (The average income after the taxes of the income recipients in the area, $R^2 = 0.3637$, $p < 0.0001$)) suggests that when KESK support is higher when the income level variables are low. There are a few outliers in every scatterplot, however, they still follow the negative trend.

The relationship of KESK support and the income level indicators show the strongest number of correlations. However, the relationship is negative. In KESK support's case, it is seen that as average taxable income, taxable median income, and average earned income goes higher, KESK support decreases. It is the same way with the taxation variables: average municipal tax paid and total average taxes are high where the KESK support is low. Somewhat obviously, the income after tax would follow the same trend. Seeing this, it can be concluded that KESK party is not the preferred party of high-income municipalities. It can also be interpreted as municipalities with lower levels of income tend to prefer KESK party over the others.

3.5 LIB (Liberaalipuolue - Vapaus valita)



The above scatterplots (in order: *Tulot* (average taxable income of the income recipients in the region in euros, $R^2 = 0.3006$, $p < 0.0001$), *Mediaanitulot* (taxable median income of the income recipients in the region in euros, $R^2 = 0.3072$, $p < 0.0001$), *Ansiotulot* (average earned income of the area's income recipients in euros, $R^2 = 0.3318$, $p < 0.0001$), *Verot* (total average taxes for income recipients in the area, $R^2 = 0.2791$, $p < 0.0001$), *Kunnallisvero* (average municipal tax paid by income recipients of the area in euros, $R^2 = 0.3081$, $p < 0.0001$), *Tulot miinus verot* (The average income after the taxes of the income recipients in the area, $R^2 = 0.3082$, $p < 0.0001$)) suggests that when the income level of a municipality is high, the LIB party support also tends to increase. There are a few outliers that still follow the trend.

In LIB party's case, the scatterplots look strange because the party support was rounded into tenths, therefore making the visualization look strange. If the decimals were not rounded, there would be more scatters on the plot. Nevertheless, the relationships are still statistically significant, as tested. It seems that average taxable income, taxable median income, and average earned income are high when the LIB party support increases (it cannot be concluded that LIB party support is high, because in fact, it is not if compared to other party supports). The taxation variables also indicate that when they are high, LIB party support increases. Moreover, the average income after taxation also indicates the same. In this case it seems that a small percentage of high-income municipalities prefer LIB party over other parties. It is important to note that only a small percentage does, since the LIB support shown is considerably little as well.

3.6 Outliers

Seen in multiple graphs in 3.2 – 3.5, there are always outliers. These outliers always have similar values of support with other municipalities, but they have higher values of income levels. These outliers have not been removed for they are true outliers – they represent the actual variation in population. After inspection, the outliers have been identified. For *Tuulonsaajia* (number of taxable income recipients in the region), the outlier is Helsinki. As the capital of Finland, it only makes sense to have high levels of taxable income recipients. Another outlier often spotted in the graphs is the area Kauniainen. Kauniainen has contrastingly high average taxable income and taxable median income, which results in high average earned income, average municipal tax, total average tax, and average income after taxations (because of the relationship between income variables).

IV. Conclusions

With the limitations raised, the result is as follows:

After analyzing income levels and municipalities' relationship on all the significantly correlated party supports, four notable parties are found with certain patterns involving the income levels. The first party, VIHR (Vihreä liitto) is found to be preferred by higher income level municipalities. This is also the same with KOK (Kansallinen Kokoomus) and LIB (Liberaalipuolue - Vapaus valita), except that LIB supporters are contrastingly little in number compared to VIHR and KOK supporters. KESK (Suomen Keskusta), on the other hand, seems to have a negative relationship with income levels – low-income level municipalities supports KESK, but higher income level municipalities do not. In short, income levels and municipalities in Finland are related to their voting behaviors to a certain degree – higher income level municipalities prefer voting for VIHR and KOK, a small percentage of them favors LIB, and the lower income level municipalities prefer to cast their votes on KESK. Thus, it is true that there are trends among certain municipalities with certain levels of income towards the municipalities' choices of politic parties in Finland (excluding Åland Islands).

In future analyses, it will be ideal to have the income and election dataset to be of the same year, and to have the same municipalities so that the entire Finland can be analyzed.