R-Bootcamp Report

Bernardo Freire Barboza da Cruz* Leonid Gavrilyuk†

20 February, 2022

*Hochschule Luzern, bernardo.dacruz@stud.hslu.ch
†Hochschule Luzern, leonid.gavrilyuk@stud.hslu.ch

Contents

1 Introduction. **3**
1.1 Project goal . . . 3
1.2 Data . . . 3

2 Data Preparation **4**
2.1 Data set 1: Turnout at the city and municipal council elections since 2006, by city district. . 4
2.2 Data set 2: Municipal elections vote share, by party and electoral district since 1913. 5
2.3 Data set 3: Wealth distribution of the population in Zürich, by district . . . 6
2.4 Data set 4: Income distribution of the population in Zürich, by district . . . 8
2.5 Combine Wealth and Income data . . . 10

3 Merging Data **12**

4 Data Visualization **13**

5 Fit Model **14**

6 Chapter of Choice TBD **15**

7 References **16**

1 Introduction.

1.1 Project goal

The goal is to understand the dynamics in the performance of each part in each electoral district of the city. The time span in interest is **2006-2018 ???** The questions to be answered: the portrait of the voters, identification of relationship (if any) between such characteristics as voters' education or age and the election results of the party.

1.2 Data

To perform analysis, six data sets covering the time period 2006-2022 were retrieved from the Open Data platform of the City of Zürich:

- Turnout at the city and municipal council elections, by city district, since 2006. (*Beteiligung am Urnengang der Stadt- und Gemeinderatswahlen nach Stadtquartier*)
- Municipal elections vote share, by party and electoral district, since 1913. (*Gemeinderatswahlen Stimmenanteil nach Partei und Wahlkreis*)
- Turnout at the city and municipal council elections, by age and gender, since 2006. ?? ??

The data sets did not have the same content and were not organised in the same way. Each data set contained some irrelevant information - for example, historical election data since 1913 whereas the time span of interest is 2006-2018. This required data preparation activities. Each data set was prepared, subsequently all of them were merged into one final table.

2 Data Preparation

2.1 Data set 1: Turnout at the city and municipal council elections since 2006, by city district.

The data set of dimensions 136x7 reflects how many people from each city district participated in five last elections (2006-2022).

Table 1: Original data set: Turnout in the city and municipal council elections since 2006, by city district

Jahr	QNr	Qname	Sberechtigte	Nteilnehmende	teilnehmende	Beteiligung
2006	11	Rathaus (Kreis 1)	1974	1186	788	39.9
2006	12	Hochschulen (Kreis 1)	377	232	145	38.5
2006	13	Lindenhof (Kreis 1)	1335	962	373	27.9
2006	14	City (Kreis 1)	597	440	157	26.3
2006	21	Wollishofen (Kreis 2)	10115	6168	3947	39.0
2006	23	Leimbach (Kreis 2)	3123	1997	1126	36.1

Examination of the dataset revealed an important issue: The territorial entity in this dataset is a city district - *"Stadtquartier"*. However, in Zürich, the elections are held across 12 electoral districts - *"Wahlkreise"*. Each electoral district consists of several "Stadtquartiers". For example, electoral district "Kreis 1+2" unites six city districts (they are shown in the "Qname" column in the Table 1 above).

To address the issue, the following manipulations were performed:

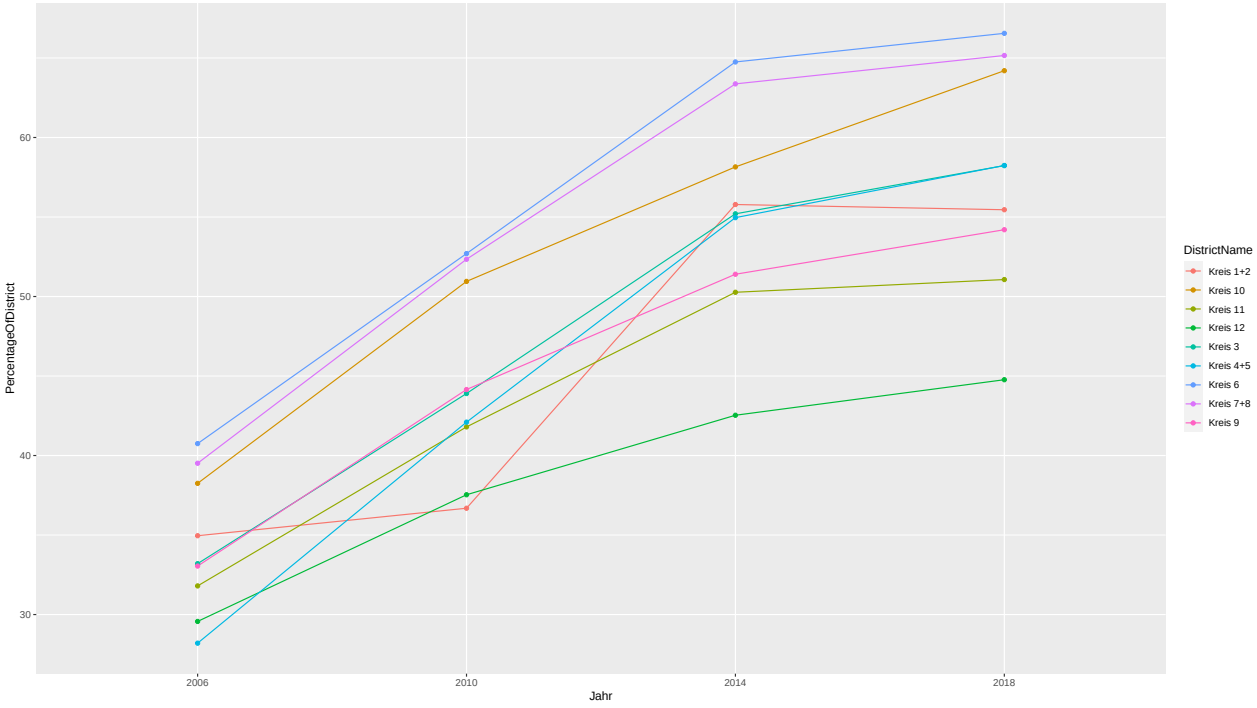
- In the column "Qname", the electoral district is specified in the parentheses: e.g. "Rathaus (Kreis 1)". The content of the parentheses was extracted with the **stringr package** and stored in the created column "DistrictName".
- Rename the electoral districts to reflect the fact that some city districts are merged for the elections - for example "Kreis 1" and "Kreis 2" became "Kreis 1+2". This was done with the **dplyr mutate() function**.
- Calculate the mean values for each group and year using **dplyr group_by() and summarise() functions** (saved as a separate data frame). The results is saved as a separate data frame, with the new column "PercentageOfDistrict".
- Calculate percentage of each district's voters relative to the total number of the city voters using **dplyr group_by() %>% summarise() %>% transmute() functions**. The results is saved as a separate data frame, with the new column "PercentageOfCity".
- Merge the dataframes with the **dplyr inner_join()**.
- Change "Jahr" and "DistrictName" to factors with the built-in **as.factor()**

Table 2: Final dataset: Turnout by electoral district

Jahr	DistrictName	PercentageOfDistrict	PercentageOfCity
2006	Kreis 1+2	34.95714	11.436022
2006	Kreis 10	38.25000	12.176240
2006	Kreis 11	31.80000	13.336140
2006	Kreis 12	29.56667	5.654176
2006	Kreis 3	33.20000	11.112771
2006	Kreis 4+5	28.20000	7.482310

The analysis shows that the amount of voters in each district has grown since 2006 - Zürich residents are becoming more active in exercising their right to vote. Kreis 6 has the highest percentage of active citizens,

followed by Kreis 7+8 and Kreis 10. Residents of Kreis 12 are the least interested in the elections people in Zürich. However, when it comes to the percentage of all city voters, Kreis 12 contributes the third largest portion of votes. This means, politicians should mobilize residents of this district to gain votes on the city and national level elections. Other important districts are Kreis 7+8 (gives most votes) and Kreis 11 (second place).



2.2 Data set 2: Municipal elections vote share, by party and electoral district since 1913.

The data set of dimensions 5100x6 shows the results of each party at the elections since 1913. Naturally, it contains a lot of irrelevant information because of the changes that happened sine 1913. For example, for each year there are separate rows for “Kreis 1 (before 2002)”, “Kreis 2 (before 2002) and”Kreis 1+2 (after 2006)" - the districts were united in 2002. Some political parties do not exist anymore; some parties changed their names. Additionally, the table is in the long format whereas the final data set requires it to be wide.

Table 3: Original data set: Municipal elections vote share, by party and electoral district since 1913.

Jahr	Partei	ParteiNr	Wahlkreis	WahlkreisSort	Stimmenanteil
1913	SP	1	Stadt Zürich	0	39.1
1913	BGB/SVP	2	Stadt Zürich	0	NA
1913	FDP	3	Stadt Zürich	0	38.6
1913	GPS	4	Stadt Zürich	0	NA
1913	GLP	5	Stadt Zürich	0	NA
1913	CVP/Die Mitte	6	Stadt Zürich	0	7.9

The following manipulations were performed:

- Choose only relevant time period (years 2006-2018) and eight still existing parties using the **%in% operator**.

- Change the names of the electoral districts. For example, "Kreis 11 (ab 1974)" became simply "Kreis 11" - the content in the parentheses was removed with the **stringr str_replace() function**.
- Rename the columns with the **dplyr %>% rename() function** to keep them in line with the other tables (e.g. "Wahlkreis" to "DistrictName").
- Remove unnecessary columns with the **dplyr %>% select() function**.
- Transform the table into wide format with the **tidyverse pivot_wider() function**.
- Merge the set with the first table "Turnout by electoral district" (see part 2.1.)

Table 4: Final dataset: Municipal elections vote share, by party and electoral district, 2006-2018.

Year	DistrictName	SP	BGB/SVP	FDP	GPS	GLP	CVP/Die Mitte	AL	EVP
2006	Kreis 1 + 2	30.1	16.2	23.1	13.1	2.4	7.7	2.5	3.0
2010	Kreis 1 + 2	28.0	16.9	19.3	13.7	8.6	5.8	2.6	1.9
2014	Kreis 1 + 2	26.6	16.2	21.0	11.4	10.5	5.1	4.9	1.9
2018	Kreis 1 + 2	30.8	13.1	19.8	12.4	10.4	4.3	6.3	1.5
2006	Kreis 3	37.5	18.2	8.6	14.3	2.3	7.1	6.1	2.3
2010	Kreis 3	34.2	16.9	8.1	13.5	10.4	5.1	6.9	1.7
2014	Kreis 3	32.1	15.0	10.5	12.8	10.4	4.0	9.8	1.4
2018	Kreis 3	35.8	10.3	10.9	14.1	10.4	3.2	12.1	1.6

The analysis shows that

2.3 Data set 3: Wealth distribution of the population in Zürich, by district

The data set of dimensions 756x8 variables reflects how the wealth distribution in absolute terms changed over time per district and per tax class. The following table shows the accumulated wealth distribution across all districts of the city of Zürich between the years 1999 and 2019.

Table 5: Original data set: Distribution wealth tax per category, district and year

SteuerJahr	KreisSort	KreisLang	SteuerTarifSort
1999	1	Kreis 1	0
1999	1	Kreis 1	1
1999	1	Kreis 1	2
1999	2	Kreis 2	0
1999	2	Kreis 2	1
1999	2	Kreis 2	2

SteuerTarifSort	SteuerTarifLang	SteuerVermoegen_p50	SteuerVermoegen_p25	SteuerVermoegen_p75
0	Grundtarif	23.0	0	174
1	Verheiratetentarif	182.0	22	711
2	Einelternfamilientarif	27.5	0	283
0	Grundtarif	37.0	3	186
1	Verheiratetentarif	148.0	33	458
2	Einelternfamilientarif	7.0	0	61

Examinating the previous showed data revealed that the data set contains a lot of irrelevant information. For example, the columns *"KreisSort"* and *"KreisLang"* are redundant, since the first in simply the encoding of

the second in number. The same applies for the the columns *"SteuerTarifSort"* and *"SteuerTarifLang"*, since the first here again, is the encoding of the second in number. However, the columns *"KreisLang"* and *"SteuerTarifLang"* are redundant and therefore, droped. For practical reasons the columns *"SteuerVermoegen_p25"* and *"SteuerVermoegen_p75"* were droped as well.
Moreover, the columns *"SteuerTarifSort"* and *"KreisSort"* are converted to factors, since those columns are automatically defined as integer number. Additionally, the names of the columns do not correspond with the previous data set and therefore, we have to change the following:

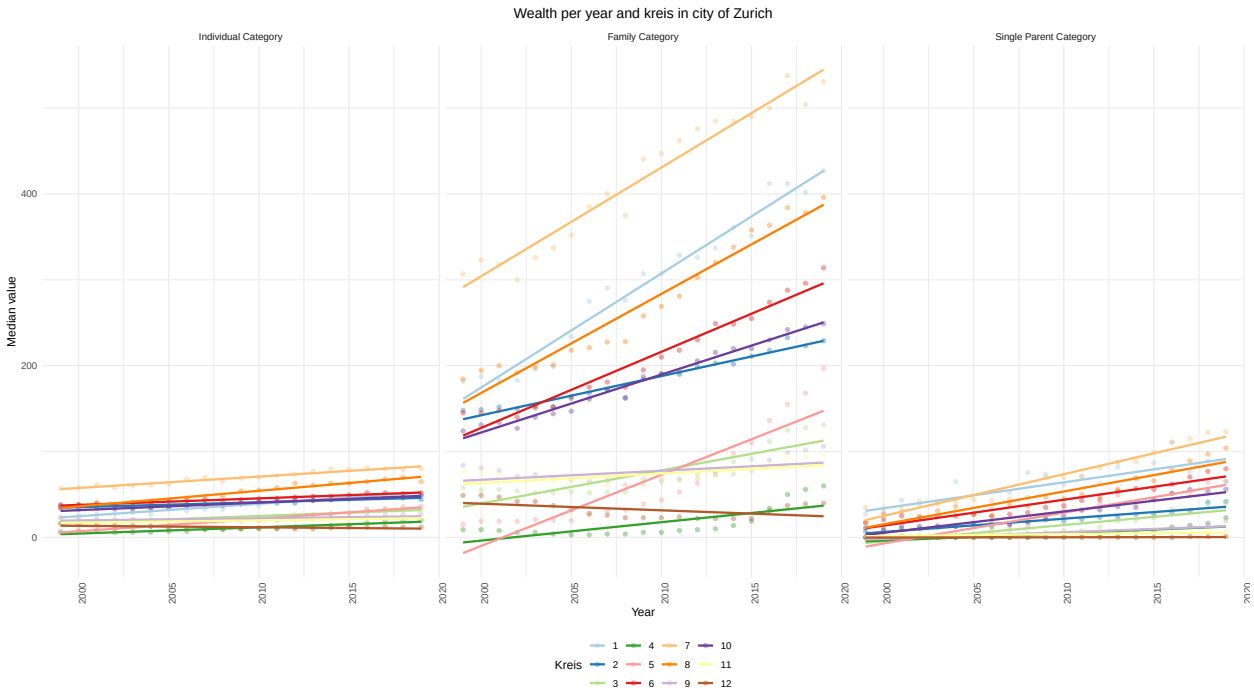
Original Column Name	New Column Name
KreisSort	DistrictNumber
SteuerJahr	Year
SteuerVermoegen_p50	Wealth
SteuerTarifSort	Category

Finally, after all modifications, the data set looks as follows:

Table 8: Final data set: Distribution wealth tax per category, district and year

Year	DistrictNumber	TaxCategory	Wealth
1999	1	0	23.0
1999	1	1	182.0
1999	1	2	27.5
1999	2	0	37.0
1999	2	1	148.0
1999	2	2	7.0

The next graph shows the distribution of wealth per year, tax category and district in Zurich.



The previous graph shows the distribution of wealth by district, year and tax category. Since the ordinate is scaled for all tax categories with the same values, it is clearly visible the difference of accumulated wealth between the districts across all and those differences seems to have a clear trend in increasing. The biggest differences between districts can be see in the *family category* subgraph. The highest values have been observed for district 7 and the lowest values for district 12. District 7, however, has the highest values among all tax categories. District 12 shows as well across all categories the lowest values of accumulated as well.

2.4 Data set 4: Income distribution of the population in Zürich, by district

The following data set of dimensions 756x8 variables reflects how the income distribution in absolute terms changed over time per district and per tax class. The following table shows the accumulated income distribution across all districts of the city of Zürich between the years 1999 and 2019.

Table 9: Original data set: Distribution income tax per category, district and year

SteuerJahr	KreisSort	KreisLang	SteuerTarifSort
1999	1	Kreis 1	0
1999	1	Kreis 1	1
1999	1	Kreis 1	2
1999	2	Kreis 2	0
1999	2	Kreis 2	1
1999	2	Kreis 2	2

SteuerTarifLang	SteuerEinkommen_p50	SteuerEinkommen_p25	SteuerEinkommen_p75
Grundtarif	37.8	17.40	64.80
Verheiratetentarif	83.4	52.00	130.20
Einelternfamilientarif	46.7	26.05	87.05
Grundtarif	37.9	19.90	58.20
Verheiratetentarif	69.7	49.10	101.40
Einelternfamilientarif	39.2	21.90	58.90

Examinating the previous showed data revealed that the data set contains a lot of irrelevant information. For example, the columns *"KreisSort"* and *"KreisLang"* are redundant, since the first in simply the encoding of the second in number. The same applies for the the columns *"SteuerTarifSort"* and *"SteuerTarifLang"*, since the first column, here again, is the encoding of the second in number. However, the columns *"KreisLang"* and *"SteuerTarifLang"* are redundant and therefore, droped. For practical reasons the columns *"SteuerEinkommen_p25"* and *"SteuerEinkommen_p75"* were droped as well.
Moreover, the columns *"SteuerTarifSort"* and *"KreisSort"* are converted to factors, since those columns are automatically defined as integer number. Additionally, the names of the columns do not correspond with the previous data set and therefore, we have to change the following:

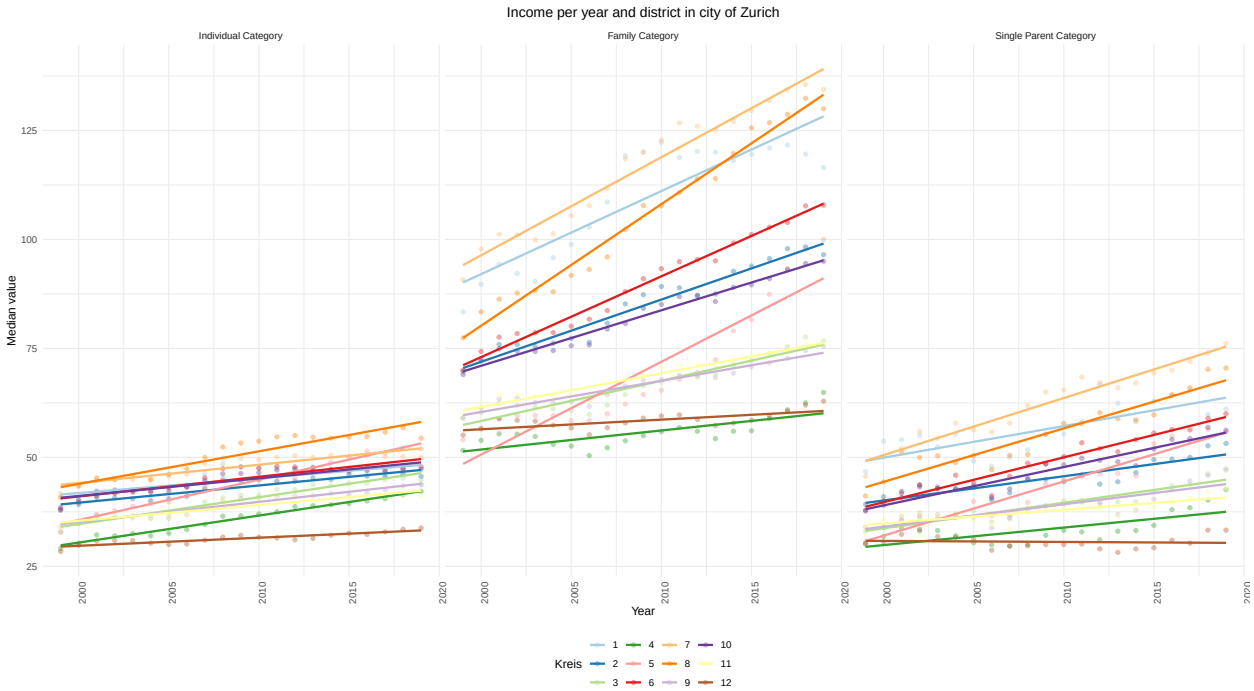
Original Column Name	New Column Name
KreisSort	DistrictNumber
SteuerJahr	Year
SteuerEinkommen_p50	Income
SteuerTarifSort	Category

Finally, after all modifications, the data set looks as follows:

Table 12: Final data set: Distribution income tax per category, district and year

Year	DistrictNumber	TaxCategory	Income
1999	1	0	37.8
1999	1	1	83.4
1999	1	2	46.7
1999	2	0	37.9
1999	2	1	69.7
1999	2	2	39.2

The next graph shows the distribution of income per year, tax category and district in Zurich.



The previous graph shows the distribution of income by district, year and tax category. Since the ordinate is scaled for all tax categories with the same values, it is clearly visible the difference of accumulated income between the districts across all and those differences seems to have a clear trend in increasing. The biggest differences between districts can be see in the *family category* subgraph. The highest values have been observed for district 7 and the lowest values for district 12. District 7, however, has the highest values among all tax categories. District 12 shows as well across all categories the lowest values of accumulated as well.

2.5 Combine Wealth and Income data

The aim of this part is to combine the income and wealth data into one data frame. As a first step, both data frames are combine using the *merge()* function. The merged data sets of income and wealth data, contains the income data and wealth data per district and tax category combined.

Table 13: Merged income and wealth data

Year	DistrictNumber	TaxCategory	Income	Wealth
1999	1	0	37.8	23.0
1999	1	1	83.4	182.0
1999	1	2	46.7	27.5
1999	10	0	38.5	34.0
1999	10	1	69.0	124.0
1999	10	2	37.7	11.0

Thereafter, the district name has to be modified in order to match if the data presented in sections concerning *Data set 1* and *Data set 2*: The name of the districts corresponds to the prefix *"Kreis"* followed by *" + "* and the corresponding number. In the case of the *wealth* and *income* data frames, those values have to be transformed and combined as the following correspondence tables shows:

Table 14: Correspondence table between district names and district numbers

Number	Name
1	Kreis 1 + 2
2	Kreis 1 + 2
3	Kreis 3
4	Kreis 4 + 5
5	Kreis 4 + 5
6	Kreis 6
7	Kreis 7 + 8
8	Kreis 7 + 8
9	Kreis 9
10	Kreis 10
11	Kreis 11
12	Kreis 12

Thereafter, several transformations have to be done as follows in order to transform the data into a mergable format:
- Pivot to a longer format the tables and combine the tax values
- Group by year, district name and tax type
- Average the values amount all categories within one district
- Pivot to wide format again to averaged values

After the step-by-step implementation of those transformation steps one gets the following data frame:

Table 15: Final wealth and income data frame per district and year

Year	DistrictName	Average Income	Average Wealth
1999	Kreis 1 + 2	52.4500	70.7500
1999	Kreis 3	41.5830	25.3330

Year	DistrictName	Average Income	Average Wealth
1999	Kreis 4 + 5	38.0165	6.1665
1999	Kreis 6	48.6670	66.6670
1999	Kreis 7 + 8	56.2580	106.0835
1999	Kreis 9	42.3830	36.0000
1999	Kreis 10	48.4000	56.3330
1999	Kreis 11	42.6000	33.0000
1999	Kreis 12	37.9330	21.6670

3 Merging Data

4 Data Visualization

5 Fit Model

6 Chapter of Choice TBD

7 References