

R-Bootcamp Report

Bernardo Freire Barboza da Cruz*

Leonid Gavriluk†

18 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	22
2.5	Combine Wealth and Income data	23
3	Merging Data	24
4	Data Visualization	25
5	Fit Model	26
6	Chapter of Choice TBD	27
7	References	28

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	Sberechtigigte	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:

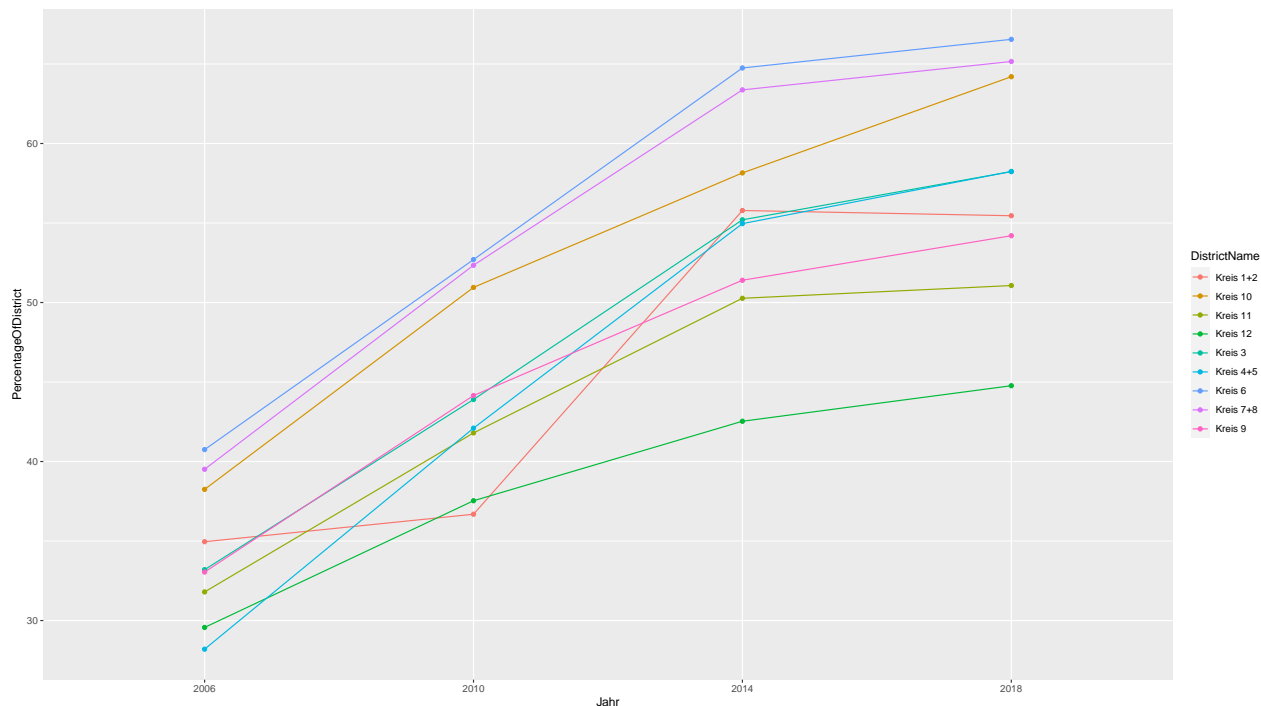
- In the column “Qname”, the electoral district is specified in the parentheses: e.g. “Rathaus (Kreis 1)”. The content of the parentheses was extracted with the **stringr** package and stored in the created column “DistrictName”.
- Rename the electoral districts to reflect the fact that some city districts 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 since 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	SteuerVermoegeen_p50	SteuerVermoegeen_p25	SteuerVermoegeen_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

Examining 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 is 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, dropped. For practical reasons the columns “*SteuerVermoegeen_p25*” and “*SteuerVermoegeen_p75*” were dropped as well.

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
SteuerVermoegeen_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	Category	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
1999	3	0	18.0
1999	3	1	58.0
1999	3	2	0.0
1999	4	0	6.0
1999	4	1	9.0
1999	4	2	0.0
1999	5	0	7.0
1999	5	1	15.0
1999	5	2	0.0
1999	6	0	38.0
1999	6	1	145.0
1999	6	2	17.0
1999	7	0	57.0
1999	7	1	307.0
1999	7	2	35.0
1999	8	0	35.0
1999	8	1	184.5
1999	8	2	18.0
1999	9	0	23.0
1999	9	1	84.0
1999	9	2	1.0
1999	10	0	34.0
1999	10	1	124.0
1999	10	2	11.0
1999	11	0	19.0
1999	11	1	77.0
1999	11	2	3.0

Year	DistrictNumber	Category	Wealth
1999	12	0	16.0
1999	12	1	49.0
1999	12	2	0.0
2000	1	0	26.0
2000	1	1	187.5
2000	1	2	30.0
2000	2	0	38.0
2000	2	1	149.0
2000	2	2	8.0
2000	3	0	19.0
2000	3	1	55.0
2000	3	2	1.0
2000	4	0	7.0
2000	4	1	9.0
2000	4	2	0.0
2000	5	0	9.0
2000	5	1	19.0
2000	5	2	0.0
2000	6	0	38.0
2000	6	1	145.0
2000	6	2	21.0
2000	7	0	58.0
2000	7	1	323.0
2000	7	2	34.0
2000	8	0	38.0
2000	8	1	194.5
2000	8	2	27.0
2000	9	0	22.0
2000	9	1	81.0
2000	9	2	2.0
2000	10	0	33.0
2000	10	1	131.0
2000	10	2	10.0
2000	11	0	18.0
2000	11	1	74.0
2000	11	2	3.0
2000	12	0	15.0
2000	12	1	49.0
2000	12	2	0.0
2001	1	0	27.0
2001	1	1	183.0
2001	1	2	43.5
2001	2	0	38.0
2001	2	1	152.0
2001	2	2	10.0
2001	3	0	20.0
2001	3	1	56.0
2001	3	2	3.0
2001	4	0	8.0
2001	4	1	8.0
2001	4	2	0.0
2001	5	0	10.0

Year	DistrictNumber	Category	Wealth
2001	5	1	19.0
2001	5	2	0.0
2001	6	0	39.0
2001	6	1	148.0
2001	6	2	25.0
2001	7	0	61.5
2001	7	1	318.0
2001	7	2	34.5
2001	8	0	40.5
2001	8	1	200.0
2001	8	2	29.5
2001	9	0	22.0
2001	9	1	78.0
2001	9	2	1.0
2001	10	0	34.0
2001	10	1	134.0
2001	10	2	13.0
2001	11	0	19.0
2001	11	1	73.0
2001	11	2	3.0
2001	12	0	16.0
2001	12	1	47.0
2001	12	2	0.5
2002	1	0	27.0
2002	1	1	183.0
2002	1	2	37.5
2002	2	0	36.0
2002	2	1	150.0
2002	2	2	10.0
2002	3	0	19.0
2002	3	1	54.0
2002	3	2	2.0
2002	4	0	6.0
2002	4	1	6.0
2002	4	2	0.0
2002	5	0	10.0
2002	5	1	19.0
2002	5	2	0.0
2002	6	0	38.0
2002	6	1	140.0
2002	6	2	23.0
2002	7	0	58.0
2002	7	1	300.0
2002	7	2	42.0
2002	8	0	38.0
2002	8	1	192.0
2002	8	2	24.5
2002	9	0	21.0
2002	9	1	71.0
2002	9	2	1.0
2002	10	0	33.0
2002	10	1	127.0

Year	DistrictNumber	Category	Wealth
2002	10	2	13.0
2002	11	0	17.0
2002	11	1	68.0
2002	11	2	3.0
2002	12	0	14.0
2002	12	1	42.0
2002	12	2	0.0
2003	1	0	29.0
2003	1	1	196.0
2003	1	2	39.0
2003	2	0	35.0
2003	2	1	153.0
2003	2	2	12.0
2003	3	0	20.0
2003	3	1	52.0
2003	3	2	2.0
2003	4	0	6.0
2003	4	1	6.0
2003	4	2	0.0
2003	5	0	10.5
2003	5	1	21.0
2003	5	2	1.0
2003	6	0	40.0
2003	6	1	151.0
2003	6	2	25.0
2003	7	0	60.0
2003	7	1	326.0
2003	7	2	41.0
2003	8	0	41.0
2003	8	1	198.5
2003	8	2	31.0
2003	9	0	21.0
2003	9	1	73.0
2003	9	2	2.0
2003	10	0	33.0
2003	10	1	140.0
2003	10	2	13.5
2003	11	0	18.0
2003	11	1	69.0
2003	11	2	2.0
2003	12	0	14.0
2003	12	1	42.0
2003	12	2	0.0
2004	1	0	27.0
2004	1	1	199.0
2004	1	2	65.0
2004	2	0	35.0
2004	2	1	152.0
2004	2	2	13.0
2004	3	0	20.0
2004	3	1	52.0
2004	3	2	2.5

Year	DistrictNumber	Category	Wealth
2004	4	0	6.0
2004	4	1	4.0
2004	4	2	0.0
2004	5	0	13.0
2004	5	1	20.0
2004	5	2	2.0
2004	6	0	41.0
2004	6	1	152.0
2004	6	2	25.0
2004	7	0	61.0
2004	7	1	337.0
2004	7	2	42.0
2004	8	0	40.0
2004	8	1	201.0
2004	8	2	36.0
2004	9	0	20.0
2004	9	1	69.0
2004	9	2	1.0
2004	10	0	34.0
2004	10	1	144.0
2004	10	2	13.0
2004	11	0	18.0
2004	11	1	66.0
2004	11	2	3.0
2004	12	0	13.0
2004	12	1	37.0
2004	12	2	0.0
2005	1	0	34.0
2005	1	1	234.0
2005	1	2	28.0
2005	2	0	37.0
2005	2	1	162.0
2005	2	2	14.0
2005	3	0	20.0
2005	3	1	53.0
2005	3	2	3.0
2005	4	0	7.0
2005	4	1	3.0
2005	4	2	0.0
2005	5	0	15.0
2005	5	1	20.0
2005	5	2	4.0
2005	6	0	43.0
2005	6	1	164.0
2005	6	2	27.0
2005	7	0	62.0
2005	7	1	352.0
2005	7	2	44.0
2005	8	0	43.0
2005	8	1	218.0
2005	8	2	27.0
2005	9	0	19.0

Year	DistrictNumber	Category	Wealth
2005	9	1	67.0
2005	9	2	1.0
2005	10	0	34.0
2005	10	1	147.0
2005	10	2	16.5
2005	11	0	17.0
2005	11	1	65.0
2005	11	2	2.0
2005	12	0	11.0
2005	12	1	32.0
2005	12	2	0.0
2006	1	0	31.0
2006	1	1	275.0
2006	1	2	36.0
2006	2	0	38.0
2006	2	1	168.0
2006	2	2	13.0
2006	3	0	20.0
2006	3	1	52.0
2006	3	2	4.0
2006	4	0	7.0
2006	4	1	3.0
2006	4	2	0.0
2006	5	0	16.0
2006	5	1	26.0
2006	5	2	9.0
2006	6	0	43.0
2006	6	1	175.0
2006	6	2	25.0
2006	7	0	68.0
2006	7	1	385.0
2006	7	2	43.5
2006	8	0	46.0
2006	8	1	221.0
2006	8	2	25.0
2006	9	0	19.0
2006	9	1	65.0
2006	9	2	1.5
2006	10	0	36.0
2006	10	1	161.0
2006	10	2	16.0
2006	11	0	18.0
2006	11	1	62.0
2006	11	2	3.0
2006	12	0	10.0
2006	12	1	28.0
2006	12	2	0.0
2007	1	0	33.0
2007	1	1	290.5
2007	1	2	52.0
2007	2	0	39.0
2007	2	1	172.5

Year	DistrictNumber	Category	Wealth
2007	2	2	16.0
2007	3	0	21.0
2007	3	1	53.0
2007	3	2	5.5
2007	4	0	9.0
2007	4	1	4.0
2007	4	2	0.0
2007	5	0	17.0
2007	5	1	31.0
2007	5	2	11.0
2007	6	0	44.0
2007	6	1	181.0
2007	6	2	27.0
2007	7	0	70.0
2007	7	1	400.0
2007	7	2	45.0
2007	8	0	48.0
2007	8	1	227.5
2007	8	2	38.5
2007	9	0	20.0
2007	9	1	66.0
2007	9	2	1.0
2007	10	0	37.0
2007	10	1	172.5
2007	10	2	15.0
2007	11	0	18.0
2007	11	1	61.0
2007	11	2	1.0
2007	12	0	10.0
2007	12	1	26.0
2007	12	2	0.0
2008	1	0	33.0
2008	1	1	276.0
2008	1	2	75.5
2008	2	0	38.0
2008	2	1	162.0
2008	2	2	18.0
2008	3	0	22.0
2008	3	1	53.0
2008	3	2	5.0
2008	4	0	9.0
2008	4	1	4.0
2008	4	2	0.0
2008	5	0	19.0
2008	5	1	32.0
2008	5	2	12.0
2008	6	0	42.0
2008	6	1	175.0
2008	6	2	29.0
2008	7	0	65.0
2008	7	1	375.0
2008	7	2	47.0

Year	DistrictNumber	Category	Wealth
2008	8	0	48.0
2008	8	1	228.0
2008	8	2	36.5
2008	9	0	20.0
2008	9	1	61.0
2008	9	2	2.0
2008	10	0	37.0
2008	10	1	163.0
2008	10	2	18.0
2008	11	0	18.0
2008	11	1	58.0
2008	11	2	2.0
2008	12	0	10.0
2008	12	1	23.0
2008	12	2	0.0
2009	1	0	39.0
2009	1	1	307.0
2009	1	2	73.5
2009	2	0	40.0
2009	2	1	184.0
2009	2	2	20.5
2009	3	0	22.0
2009	3	1	62.0
2009	3	2	6.0
2009	4	0	10.0
2009	4	1	6.0
2009	4	2	1.0
2009	5	0	19.0
2009	5	1	39.0
2009	5	2	27.0
2009	6	0	44.0
2009	6	1	195.0
2009	6	2	35.0
2009	7	0	69.0
2009	7	1	440.5
2009	7	2	70.0
2009	8	0	54.0
2009	8	1	258.0
2009	8	2	47.0
2009	9	0	21.0
2009	9	1	68.0
2009	9	2	3.0
2009	10	0	39.0
2009	10	1	187.0
2009	10	2	23.0
2009	11	0	19.0
2009	11	1	66.0
2009	11	2	3.0
2009	12	0	10.0
2009	12	1	23.0
2009	12	2	0.0
2010	1	0	37.0

Year	DistrictNumber	Category	Wealth
2010	1	1	307.0
2010	1	2	61.5
2010	2	0	41.0
2010	2	1	190.0
2010	2	2	19.0
2010	3	0	24.0
2010	3	1	65.0
2010	3	2	8.0
2010	4	0	10.0
2010	4	1	6.0
2010	4	2	1.0
2010	5	0	20.0
2010	5	1	43.5
2010	5	2	32.5
2010	6	0	45.0
2010	6	1	210.0
2010	6	2	37.0
2010	7	0	70.0
2010	7	1	447.0
2010	7	2	70.0
2010	8	0	55.0
2010	8	1	269.0
2010	8	2	45.0
2010	9	0	21.0
2010	9	1	68.0
2010	9	2	4.0
2010	10	0	41.0
2010	10	1	191.0
2010	10	2	25.0
2010	11	0	19.0
2010	11	1	68.0
2010	11	2	2.0
2010	12	0	11.0
2010	12	1	23.0
2010	12	2	0.0
2011	1	0	46.0
2011	1	1	328.5
2011	1	2	71.5
2011	2	0	40.0
2011	2	1	190.0
2011	2	2	21.0
2011	3	0	25.0
2011	3	1	68.0
2011	3	2	8.0
2011	4	0	11.0
2011	4	1	8.0
2011	4	2	2.0
2011	5	0	20.0
2011	5	1	53.0
2011	5	2	31.0
2011	6	0	45.0
2011	6	1	218.0

Year	DistrictNumber	Category	Wealth
2011	6	2	43.0
2011	7	0	72.0
2011	7	1	462.0
2011	7	2	70.5
2011	8	0	58.0
2011	8	1	281.0
2011	8	2	50.0
2011	9	0	21.0
2011	9	1	67.0
2011	9	2	5.0
2011	10	0	42.0
2011	10	1	193.0
2011	10	2	32.0
2011	11	0	20.0
2011	11	1	66.0
2011	11	2	3.0
2011	12	0	10.0
2011	12	1	24.0
2011	12	2	0.0
2012	1	0	46.0
2012	1	1	326.0
2012	1	2	63.0
2012	2	0	40.0
2012	2	1	199.0
2012	2	2	23.5
2012	3	0	26.0
2012	3	1	74.0
2012	3	2	12.0
2012	4	0	12.0
2012	4	1	9.0
2012	4	2	2.0
2012	5	0	23.0
2012	5	1	63.0
2012	5	2	39.0
2012	6	0	47.0
2012	6	1	230.0
2012	6	2	44.0
2012	7	0	73.0
2012	7	1	476.0
2012	7	2	81.0
2012	8	0	63.0
2012	8	1	303.0
2012	8	2	48.0
2012	9	0	21.0
2012	9	1	68.5
2012	9	2	4.0
2012	10	0	44.0
2012	10	1	205.5
2012	10	2	32.0
2012	11	0	21.0
2012	11	1	70.0
2012	11	2	3.0

Year	DistrictNumber	Category	Wealth
2012	12	0	10.0
2012	12	1	22.0
2012	12	2	0.0
2013	1	0	47.0
2013	1	1	337.0
2013	1	2	82.0
2013	2	0	43.0
2013	2	1	203.0
2013	2	2	24.0
2013	3	0	27.0
2013	3	1	77.0
2013	3	2	19.0
2013	4	0	13.0
2013	4	1	10.0
2013	4	2	4.0
2013	5	0	26.0
2013	5	1	73.5
2013	5	2	49.0
2013	6	0	48.0
2013	6	1	249.0
2013	6	2	54.0
2013	7	0	77.0
2013	7	1	485.0
2013	7	2	87.5
2013	8	0	61.0
2013	8	1	320.0
2013	8	2	56.0
2013	9	0	21.0
2013	9	1	73.0
2013	9	2	5.0
2013	10	0	43.0
2013	10	1	215.5
2013	10	2	36.0
2013	11	0	21.0
2013	11	1	72.5
2013	11	2	3.0
2013	12	0	10.0
2013	12	1	22.0
2013	12	2	0.0
2014	1	0	45.0
2014	1	1	361.0
2014	1	2	70.0
2014	2	0	44.0
2014	2	1	202.0
2014	2	2	27.0
2014	3	0	28.0
2014	3	1	88.0
2014	3	2	20.0
2014	4	0	14.0
2014	4	1	14.0
2014	4	2	5.0
2014	5	0	26.0

Year	DistrictNumber	Category	Wealth
2014	5	1	98.0
2014	5	2	39.0
2014	6	0	48.0
2014	6	1	248.5
2014	6	2	57.0
2014	7	0	80.0
2014	7	1	484.5
2014	7	2	88.5
2014	8	0	63.0
2014	8	1	338.0
2014	8	2	68.5
2014	9	0	23.0
2014	9	1	74.0
2014	9	2	7.0
2014	10	0	44.0
2014	10	1	220.0
2014	10	2	35.0
2014	11	0	21.0
2014	11	1	75.0
2014	11	2	4.0
2014	12	0	11.0
2014	12	1	22.0
2014	12	2	0.0
2015	1	0	50.0
2015	1	1	351.0
2015	1	2	87.0
2015	2	0	45.0
2015	2	1	211.0
2015	2	2	27.0
2015	3	0	29.0
2015	3	1	91.5
2015	3	2	23.0
2015	4	0	14.5
2015	4	1	19.0
2015	4	2	6.0
2015	5	0	28.0
2015	5	1	110.5
2015	5	2	43.5
2015	6	0	49.0
2015	6	1	255.0
2015	6	2	56.0
2015	7	0	79.0
2015	7	1	490.0
2015	7	2	99.0
2015	8	0	63.0
2015	8	1	358.0
2015	8	2	66.0
2015	9	0	23.0
2015	9	1	78.0
2015	9	2	7.0
2015	10	0	45.0
2015	10	1	220.0

Year	DistrictNumber	Category	Wealth
2015	10	2	44.0
2015	11	0	22.0
2015	11	1	78.0
2015	11	2	5.0
2015	12	0	12.0
2015	12	1	22.0
2015	12	2	0.0
2016	1	0	49.0
2016	1	1	412.0
2016	1	2	111.0
2016	2	0	48.0
2016	2	1	218.0
2016	2	2	29.0
2016	3	0	31.0
2016	3	1	112.0
2016	3	2	33.0
2016	4	0	17.0
2016	4	1	34.0
2016	4	2	12.0
2016	5	0	33.0
2016	5	1	136.0
2016	5	2	51.0
2016	6	0	52.0
2016	6	1	274.0
2016	6	2	62.0
2016	7	0	81.0
2016	7	1	500.0
2016	7	2	111.0
2016	8	0	65.0
2016	8	1	363.5
2016	8	2	80.0
2016	9	0	26.0
2016	9	1	90.0
2016	9	2	12.0
2016	10	0	47.0
2016	10	1	230.0
2016	10	2	51.5
2016	11	0	22.0
2016	11	1	85.0
2016	11	2	6.0
2016	12	0	12.0
2016	12	1	31.0
2016	12	2	0.0
2017	1	0	47.0
2017	1	1	412.0
2017	1	2	77.0
2017	2	0	47.0
2017	2	1	232.5
2017	2	2	31.0
2017	3	0	32.0
2017	3	1	125.0
2017	3	2	31.0

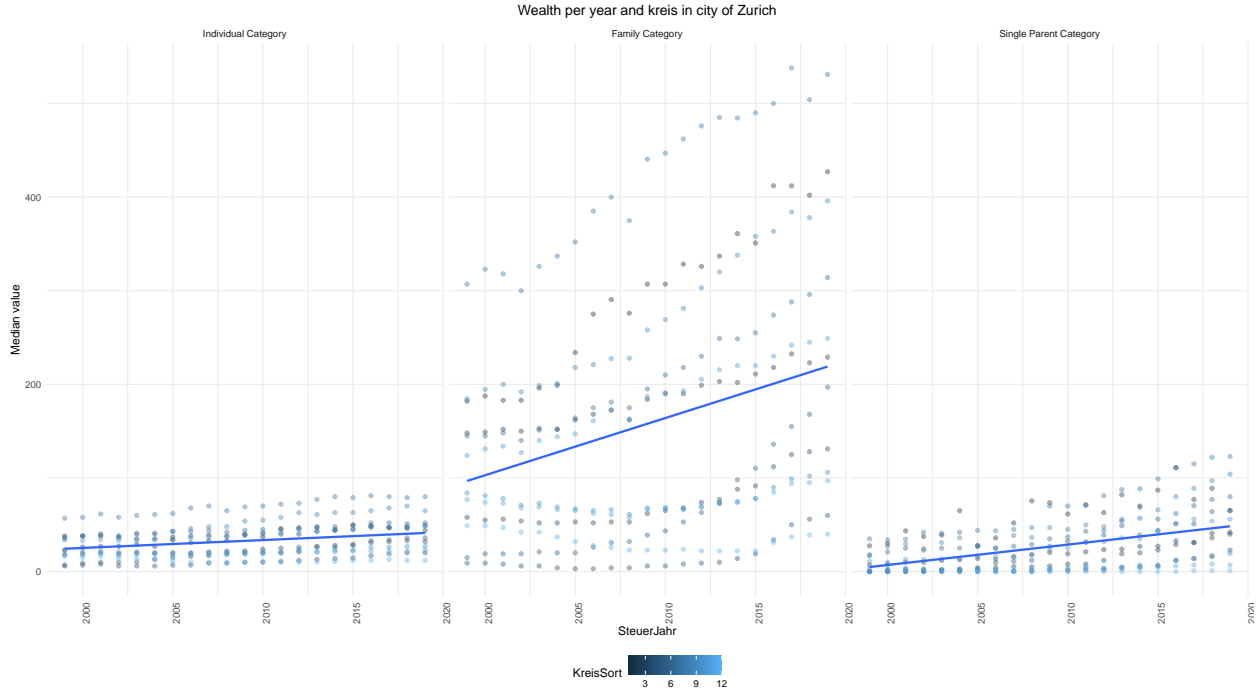
Year	DistrictNumber	Category	Wealth
2017	4	0	19.0
2017	4	1	50.0
2017	4	2	14.0
2017	5	0	34.0
2017	5	1	155.0
2017	5	2	51.0
2017	6	0	52.0
2017	6	1	288.0
2017	6	2	70.5
2017	7	0	80.0
2017	7	1	538.0
2017	7	2	115.0
2017	8	0	68.0
2017	8	1	384.0
2017	8	2	89.0
2017	9	0	26.0
2017	9	1	99.0
2017	9	2	14.0
2017	10	0	47.0
2017	10	1	242.0
2017	10	2	56.0
2017	11	0	22.0
2017	11	1	94.0
2017	11	2	7.0
2017	12	0	13.0
2017	12	1	37.0
2017	12	2	1.0
2018	1	0	48.0
2018	1	1	402.0
2018	1	2	89.0
2018	2	0	46.0
2018	2	1	223.0
2018	2	2	41.0
2018	3	0	33.0
2018	3	1	128.0
2018	3	2	36.0
2018	4	0	20.0
2018	4	1	56.0
2018	4	2	16.0
2018	5	0	36.0
2018	5	1	168.0
2018	5	2	64.0
2018	6	0	51.0
2018	6	1	296.0
2018	6	2	77.0
2018	7	0	79.0
2018	7	1	504.0
2018	7	2	122.0
2018	8	0	70.0
2018	8	1	378.0
2018	8	2	97.0
2018	9	0	27.0

Year	DistrictNumber	Category	Wealth
2018	9	1	102.0
2018	9	2	17.0
2018	10	0	47.0
2018	10	1	245.0
2018	10	2	54.0
2018	11	0	22.0
2018	11	1	95.0
2018	11	2	8.0
2018	12	0	12.0
2018	12	1	39.0
2018	12	2	1.0
2019	1	0	49.0
2019	1	1	427.0
2019	1	2	65.0
2019	2	0	44.0
2019	2	1	229.0
2019	2	2	42.0
2019	3	0	32.0
2019	3	1	131.0
2019	3	2	40.0
2019	4	0	20.0
2019	4	1	60.0
2019	4	2	23.0
2019	5	0	36.0
2019	5	1	197.0
2019	5	2	65.5
2019	6	0	52.0
2019	6	1	314.0
2019	6	2	80.0
2019	7	0	80.0
2019	7	1	531.0
2019	7	2	123.0
2019	8	0	65.0
2019	8	1	396.0
2019	8	2	104.0
2019	9	0	27.0
2019	9	1	106.0
2019	9	2	19.5
2019	10	0	49.0
2019	10	1	249.0
2019	10	2	56.0
2019	11	0	23.0
2019	11	1	97.0
2019	11	2	7.0
2019	12	0	12.0
2019	12	1	40.0
2019	12	2	1.0

SteuerJahr	KreisSort	SteuerTarifSort	SteuerVermoegeen_p50
1999	1	0	23.0

SteuerJahr	KreisSort	SteuerTarifSort	SteuerVermoegeen_p50
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.



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

SteuerJahr	KreisSort	KreisLang	SteuerTarifSort	SteuerTarifLang	SteuerEinkommenSteuer	EinkommenSteuer	EinkommenSteuer	Einkommen_p75
1999	1	Kreis 1	0	Grundtarif	37.8	17.40	64.80	
1999	1	Kreis 1	1	Verheiratetentarif	83.4	52.00	130.20	
1999	1	Kreis 1	2	Einelternfamilientarif	46.7	26.05	87.05	
1999	2	Kreis 2	0	Grundtarif	37.9	19.90	58.20	
1999	2	Kreis 2	1	Verheiratetentarif	69.7	49.10	101.40	
1999	2	Kreis 2	2	Einelternfamilientarif	39.2	21.90	58.90	

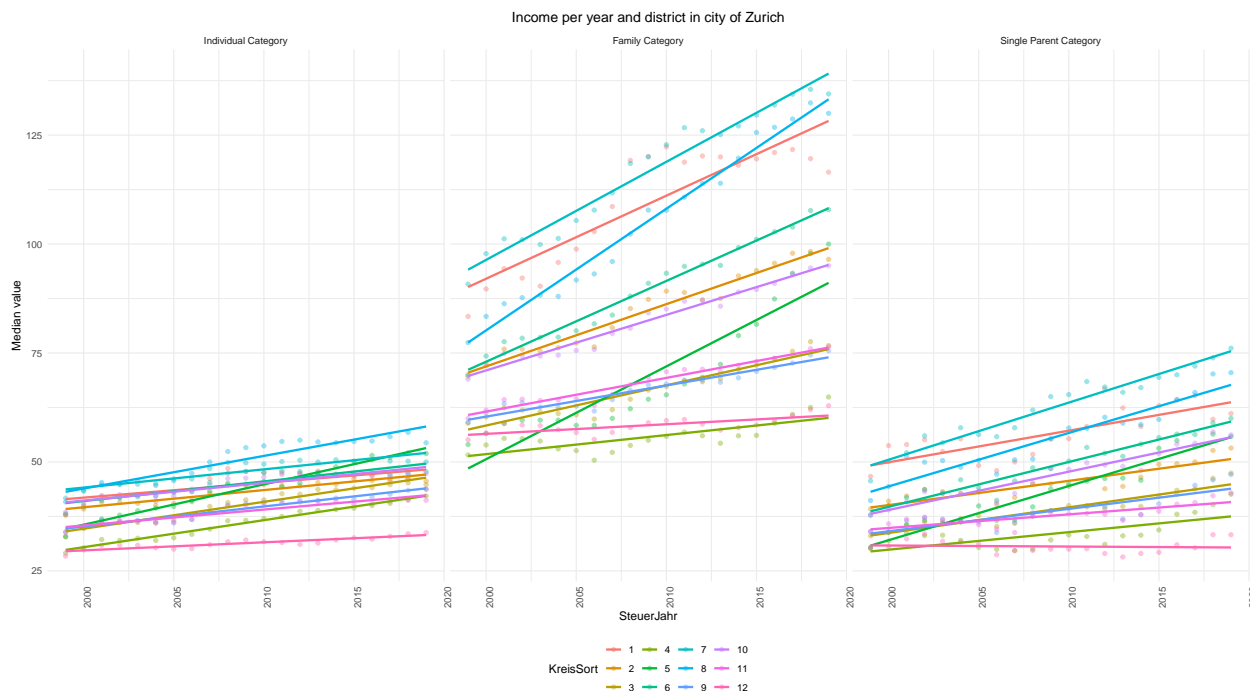
Examining the previous showed data revealed that the data set contains a lot of irrelevant information. For example, the columns “KreisSort” and “KreisLang” are redundant, since the first is simply the encoding of

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 not redundant and therefore, dropped.

Table 11: Final data set: Distribution wealth tax per category, district and year

SteuerJahr	KreisSort	SteuerTarifSort	SteuerEinkommen_p50	SteuerEinkommen_p25	SteuerEinkommen_p75
1999	1	0	37.8	17.40	64.80
1999	1	1	83.4	52.00	130.20
1999	1	2	46.7	26.05	87.05
1999	2	0	37.9	19.90	58.20
1999	2	1	69.7	49.10	101.40
1999	2	2	39.2	21.90	58.90

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



2.5 Combine Wealth and Income data

The aim of this part is to combine the income and wealth data into one dataframe.

Table 12: Merged income and wealth data

SteuerJahr	KreisSort	SteuerTarifSort	SteuerEinkommen_p50	SteuerEinkommen_p25	SteuerEinkommen_p75	SteuerVermoegeen_p50
1999	1	0	37.8	17.40	64.80	23.0
1999	1	1	83.4	52.00	130.20	182.0
1999	1	2	46.7	26.05	87.05	27.5
1999	10	0	38.5	20.00	57.30	34.0
1999	10	1	69.0	49.60	98.40	124.0
1999	10	2	37.7	20.65	54.15	11.0

3 Merging Data

4 Data Visualization

5 Fit Model

6 Chapter of Choice TBD

7 References