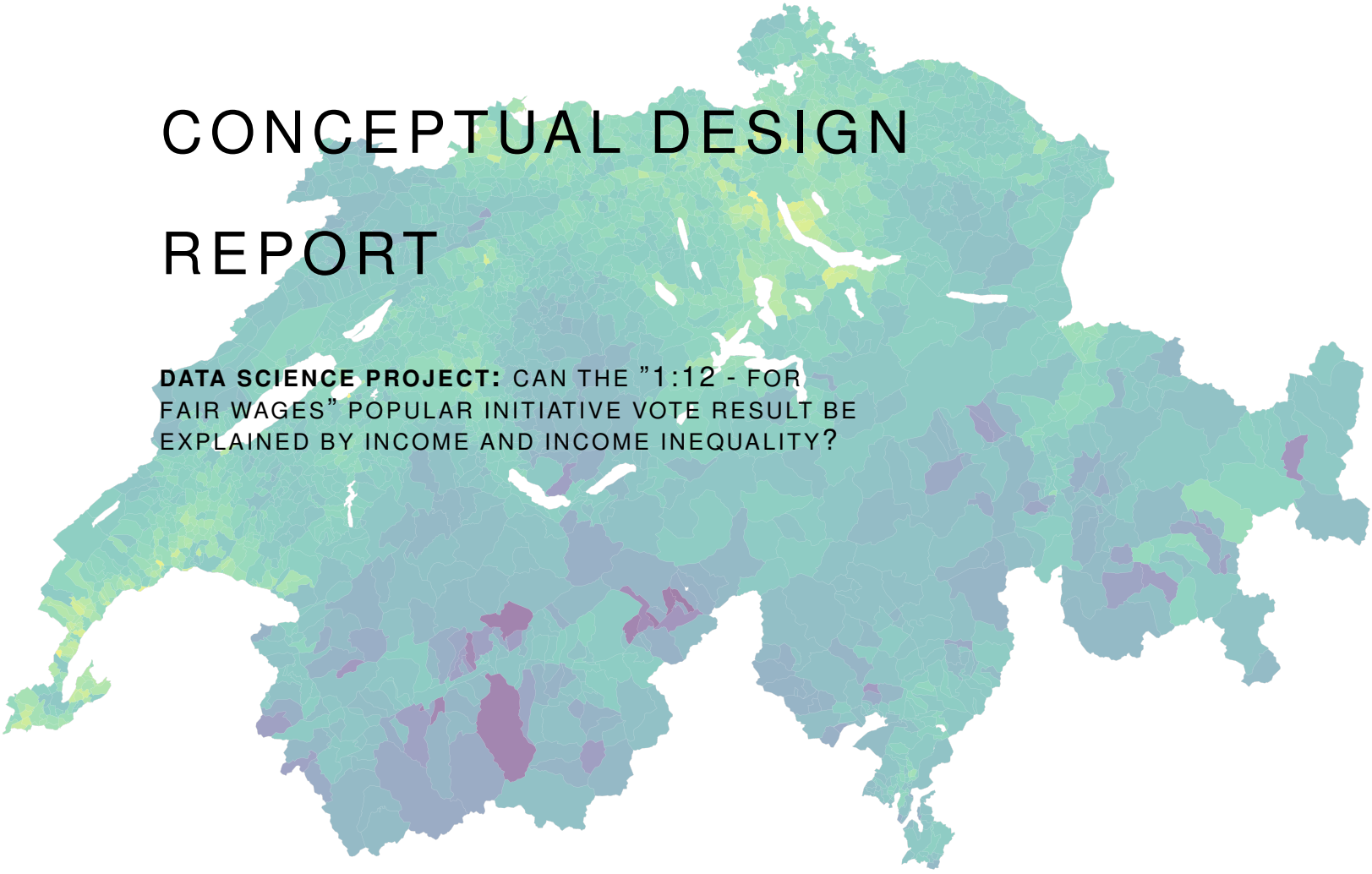
Petra Müller

CONCEPTUAL DESIGN REPORT

DATA SCIENCE PROJECT: CAN THE ”1:12 - FOR FAIR WAGES” POPULAR INITIATIVE VOTE RESULT BE EXPLAINED BY INCOME AND INCOME INEQUALITY?



CAS Applied Data Science

Author contact information: Petra Müller, Holenackerstrasse 65, 3027 Bern

E-Mail: petra_mueller@students.unibe.ch

GitHub: https://github.com/pmuellerCAS

This document was created in LaTeXusing the Tufte LaTeX package, A Tufte-inspired LaTeX class for producing handouts, papers, and books by Kevin Godby, Bil Kleb, and Bill Wood published on GitHub (https://github.com/Tufte-LaTeX/tufte-latex) licensed under the Apache License 2.0 (https://www.apache.org/licenses/LICENSE-2.0).

27. September 2019

Abstract

In 2013, Swiss citizens voted on a popular initiative - ”1:12 - for fair wages” on federal level demanding to limit the amount of executive pay within a company to a maximum of 12 times that of the lowest paid workers. The popular initiative was rejected to a great extent. But why did the Swiss populace turn down this opportunity to reduce the wage gap? This question is addressed partially by this Data Science Project. Correlations between income, income inequality (Gini-coefficient) and vote result were examined. Gini-coefficient and vote result were not correlated (pearson, r=-.056). Median income and acceptance of the Popular Initiative were slightly negatively correlated with r=-.39. Although the correlation between median income and vote result is slight at most and moderated by canton, it nonetheless indicates an interesting relationship. But correlation does not equal causation and no explanation for this relationship can be provided at this point, thus further analyses are needed to investigate the relationship between income and voting behaviour.

Contents

Objectives

Primary objectives

The main goal of this data science project was to examine the relationship between the acceptance of the Popular Initiative "1:12 - for fair wages" in Switzerland in the year 2013 and income as well as income inequality on a communal level. The following research questions were stated:

- Is there a relationship between income inequality (Gini-coefficient) and acceptance of the ”1:12 - for fair wages” Popular Initiative?
- Is there a relationship between wealth (income) and acceptance of the ”1:12 - for fair wages” Popular Initiative?

Thus the analysis objective is to uncover possible correlations between income, income inequality and the vote result as well as to examine possible correlations more detailed to maybe uncover other factors (moderators) that might shape the relationship between variables.

Secondary objectives

Out of personal curiosity and interest the secondary objective of this project is to visualise voting data as well as income data by use of geographical maps, that is to combine geospatial data (polygons) and vote/tax data to create informative geographical map plots.

Methods

Infrastructure

The hardware used for this project is a
HP ProBook 470 G2 using a **Intel© Core™ i7-4510U CPU @ 2.00GHz × 2** processor and 15.5 GiB of memory.

The operating system is
Manjaro Linux (Kernel: 5.1.21-1-MANJARO) with Cinnamon Desktop Environment (version 4.2.4).

Software, environments, packages and libraries

To ease data collection, the software **LibreOffice Calc** version 6.3.0.4 was partially used to view and preprocess data. The data science distribution **Anaconda 3** (Anaconda3-2019.07-Linux-x86_64) is used to install and maintain development environments, libraries and packages for R and Python. The development environment is **JupyterLab** version 1.1.3 running Python version 3.7.3 and R version 3.5.1 "Feather Spray".

Packages and Libraries: The following R packages were used in this project: readr, dplyr, tidyr. The following Python libraries were used in this project: pandas, numpy, scipy, sklearn, geopandas, matplotlib, seaborn, bokeh, pandas_bokeh

geopandas is a library that facilitates working with geospatial data in Python by extending the datatypes used by pandas to allow spatial operations on geometric types (http://geopandas.org/). **seaborn** is a Python data visualization library based on matplotlib (https://seaborn.pydata.org/) and **bokeh** is an interactive visualization library that targets modern web browsers for presentation (https://bokeh.pydata.org) - it was used to produce interactive map plots.

Statistical methods

Two main statistical methods are applied in the analysis, namely the calculation of **pearson correlations** and **linear regression** to assess strength and direction of the association between variables. **Descriptive statistics** are used to analyse and visualise variability and distribution of variables.

Data

Voting data

Collecting voting data for the "1:12 - for fair wages" Popular Initiative turned out to be challenging, as the vote took place in 2013 but the Federal Statistical Office FSO offers vote results on communal level only for the last five years. For older vote results, one has to consult archived data[1] - which introduces the problem that the archived data does not contain the communes in their historical (2013) state, but already applied all commune fusions up to 01.04.2018 which would result in a loss of more than 70 communes[2]. Therefore, the archived data from the FSO could only be used for some cantons, namely OW, NW, AI, AR, BL, BS, UR, SZ, ZG and TG. To obtain voting data for the remaining cantons (AG, FR, LU, GL, JU, NE, SG, SH, SO, TI, VD, VS, GR, BE and ZH), voting data was downloaded directly from the archives of the individual State Chancelleries[3]. Because the individual State Chancelleries used a variety of data formats (xlsx, csv, pdf, txt, etc) and content, LibreOffice Calc as well as R was used to manually preprocess and combine these files into one single sheet containing **commune name, turnout rate, number of (valid) 'yes' votes, number of (valid) 'no' votes** as well as the communes **canton abbrevation.** This resulted in a dataset containing N = 2385 communes.

By the time this report was written, a more convenient way to obtain historical vote data was found, namely the data provided by the Federal Statistical Office FSO in the 'Political Atlas of Switzerland': `https://www.atlas.bfs.admin.ch/maps/12/de/9066_8684_5401_259/16694.html`. **I strongly recommend using this data source**.

[1] from `https://www.pxweb.bfs.admin.ch/pxweb/de/px-x-1703030000_101/px-x-1703030000_101/px-x-1703030000_101.px`

[2] see `https://de.wikipedia.org/wiki/Gemeindefusionen_in_der_Schweiz`

[3] For a full list of links please see data_sources.txt

Tax data

Tax (Income) data was obtained from the Federal Tax Administration[4]. The 2014 publication was used because it contains the tax data that was collected in 2013. It contains fewer communes (N = 2352) than the voting data because it portrays the communes in their 2014 state. Again, LibreOffice Calc was used to preprocess the data by deleting irrelevant columns. The therefore preprocessed file contained **commune name, canton name, commune number (bfsnr),**

[4] Download link: `www.estv2.admin.ch/d/dokumentation/zahlen_fakten/karten/2014/equivalent/Daten.xls`

number of taxed households, mean equivalised income, median equivalised income and **Gini-Coefficient** per commune[5].

[5] For an explanation of these variables see Chapter 'Metadata'.

Geospatial data

GEOGRAPHICAL DATA is used to visualise the variables and analysis results on map plots. The Federal Statistical Office FSO as well as the Federal Office of Topography swisstopo both offer a variety of geospatial data. In this project, generalised commune shapes in their 2014 state[6] were used to match tax/income data. This geospatial data comes in the form of so called shapefiles[7], that can be read in Python using the Geopandas library. For this project, the file **'g1g14vz'** was used, which contains N = 2356 generalised geometry shapes of communes[8] as they were on the 31.12.2014, listing **commune name, commune number, canton number, district number, greater region number, size (in hectares)** per commune as well as a **variety of geospatial information**[9]**.**

[6] Downloaded from the FSO (GEOSTAT): `https://www.bfs.admin.ch/bfsstatic/dam/assets/328824/master`

[7] https://en.wikipedia.org/wiki/Shapefile

[8] To be precise, 2352 communes and 4 special regions that either belong to multiple communes or no commune.

[9] see Chapter 'Metadata'

Analysis dataset

VOTE AND INCOME/TAX DATA were combined into one single dataframe. For this to be possible, a few adjustments had to be made, namely:

- **Vote data:** 2014 commune fusions were implemented by aggregating the affected communes[10] into new ones (sum of 'yes' and 'no' votes, mean of turnout rate, new commune name) to match tax/income data. Some communes were manually re-named to match tax/income data[11].
- **Tax data:** The canton of Bern (BE) collects the votes of some communes[12] pooled, thus the income/tax data of the affected communes had to be aggregated (sum of taxed households, mean of mean equivalised income, mean of median equivalised income, mean Gini-Coefficient) to match vote data.

This resulted in a single dataframe[13] containing tax as well as vote information for 2343 communes. For reproducibility reasons, all the raw data files are provided[14].

[10] For a list of all the affected communes see 'commune_fusions.txt'

[11] This was done manually in LibreOffice Calc after comparing commune names in R. It usually implied adding a whitespace after a dot, e.g. 'St.Martin'–>'St. Martin', adding the canton abbrevation after the name, e.g. 'Wald' –> 'Wald (BE)' or to spell abbrevations in full, e.g. 'Sils i.D.' –> 'Sils im Domleschg'.

[12] For a list of all the affected communes, see 'pooled_communes.txt'

[13] See 'analysis_dataset.csv'

[14] See 'Raw_data' folder and data_sources.txt

Data visualisation

Tables and graphs:

The follwing tables and graphs portray the analysis dataset and the relevant variables that will be used in this project.

	commune	vote_yes	vote_no	turnout_perc	canton	canton_name	bfsnr	n_taxed	mean_income	median_income	gini-coeff	percent_yes
0	Mulegns	1	9	37.04	GR	Graubünden / Grigioni / Grisch	3534	16	37570.14	30422.22	0.31	10.00
1	Urmein	8	60	60.18	GR	Graubünden / Grigioni / Grisch	3670	99	39120.33	34500.00	0.47	11.76
2	Autafond	5	31	67.92	FR	Fribourg / Freiburg	2172	37	39946.14	38600.00	0.37	13.89
3	Greng	12	72	66.67	FR	Fribourg / Freiburg	2261	112	115748.89	67188.89	0.59	14.29
4	Flims	134	801	52.49	GR	Graubünden / Grigioni / Grisch	3732	2273	39006.96	29300.00	0.51	14.33

Table 1: The first 5 rows of the analysis dataset. Numbers are rounded to 2 decimals.

	vote_yes	vote_no	turnout_perc	bfsnr	n_taxed	mean_income	median_income	gini-coeff	percent_yes
count	2343.00	2343.00	2343.00	2343.00	2343.00	2343.00	2343.00	2343.00	2343.00
mean	401.92	760.74	55.71	3324.49	2129.59	47369.02	41169.82	0.40	34.16
std	1426.89	2026.73	7.78	2117.14	7453.23	26138.46	7408.75	0.07	8.51
min	0.00	0.00	0.00	0.00	16.00	17140.96	4800.00	0.23	0.00
25%	79.00	149.00	50.86	1124.00	366.00	40023.35	37333.33	0.35	28.70
50%	170.00	339.00	54.89	3424.00	822.00	44879.45	41387.18	0.38	32.92
75%	376.50	785.00	60.10	5405.50	2000.50	50251.60	45469.44	0.42	38.33
max	47878.00	68251.00	97.14	6810.00	254158.00	1021727.35	68600.00	0.97	79.41

Table 2: Analysis dataset summary. Numbers are rounded to 2 decimals.

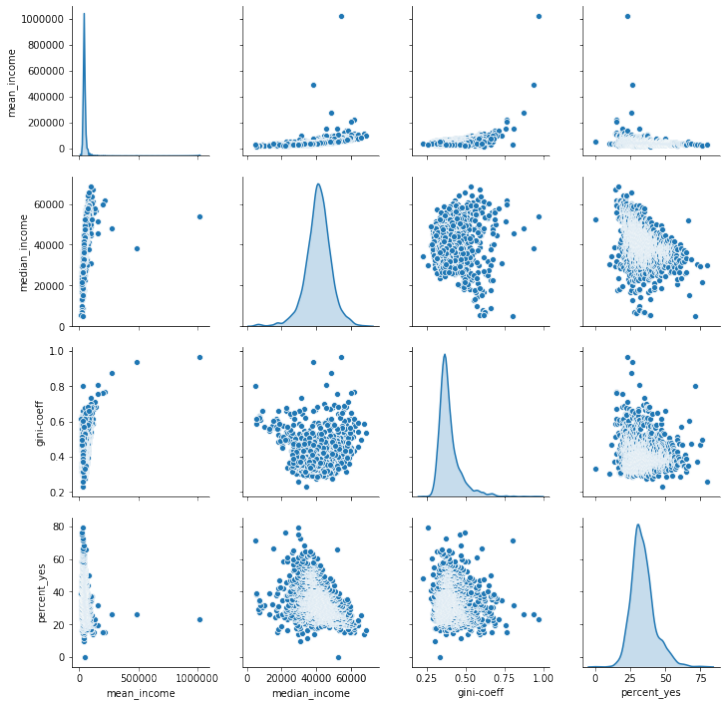


Figure 1: Pairplot of the analysis variables created with Seaborn.

Geographical map plots:

To produce informative map plots, geospatial data had to be merged to form new polygons for the pooled communes in the canton BE, similar to tax data aggregation. This was done in Python by forming unions of the affected communes polygons[15].

[15] For an in-depth explanation see Pmueller_project_analysis.jpynb

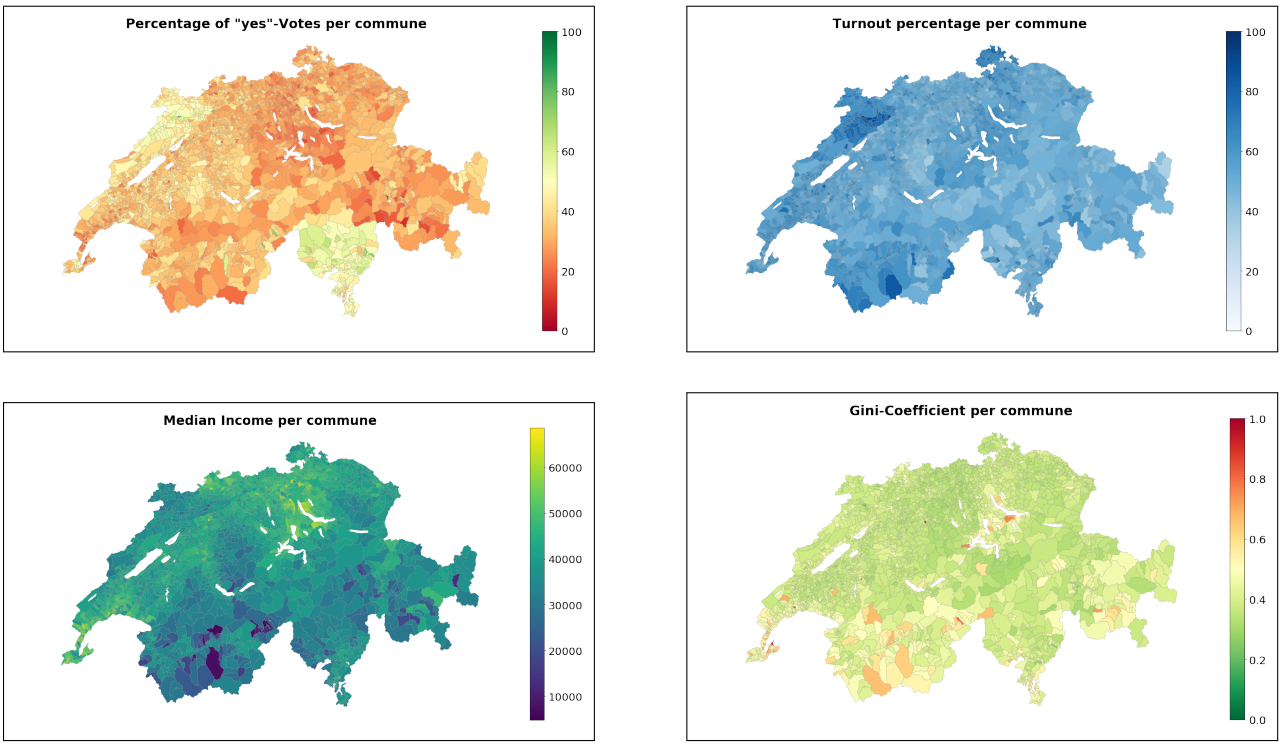


Figure 2: Geographical map plots created with GeoPandas.

Interactive plots:

Similar geographical map plots can also be rendered using the Pandas_Bokeh and the Bokeh Python libraries but with the advantage of interactivity, e.g. hover tools and zooming. An example is illustrated in the Jupyter Notebook (Pmueller_project_analysis.ipynb) and the output can also be found in the "Publications" folder ("Interactive Plot.html").

Metadata

Explanations of the variables in the analysis dataset and the geospatial data are available in the 'Metadata' folder, namely

- 'analysis_dataset_variables.txt': A text file containing descriptions of the analysis dataset variables
- 'geodata_variables.txt': A text file containing descriptions of the geographical data variables
- 'geodata_explanation.pdf: The official pdf document provided by the Federal Statistical Office FSO containing an in-depth explanation of the geographical data files and variables (in german)
- 'data_sources.txt': A text file containing all the data sources (links) to the data used in this project as they were at the time of data collection

Data Quality

Neither the collected vote data nor the tax data contained any missing values in the variables needed to conduct the analysis, therefore no further data cleaning was required after data pre-processing[16]. Nonetheless, there remains one issue that has to be addressed when assessing vote data quality, that is the question of how to handle votes of Swiss citizens living abroad.

[16] Cleaning and aggregating of raw data is described in chapter 'Data'.

Swiss citizens voting from abroad are neither living in the commune they voted in nor are they taxed in that commune, thus to compare voting behavior and income on a communal level these votes had to be excluded from the analysis. This however was not possible for all Cantons.

The problem is that not all Cantons report votes from Swiss Expats in the same way. The Canton Jura (JU) for example reports how many Swiss Expats were eligible to vote per commune, but not the actual number or result of the Expat's votes - therefore these votes could not be excluded. The Canton Schaffhausen (SH) does not mention Expats voting at all, thus it could be assumed they are already included in the votes per commune, inseparable. The Canton Fribourg (FR) then reports Expats votes on the Cantonal level as if they form a separate district. In this case, the Expat's votes were not included in the analysis dataset.

Therefore for the Cantons where Expat's votes were inseparable from other votes per commune, the Expat's votes were included in the analysis. This affected Cantons ZH, BE, SZ, OW, NW, GL, ZG, SO, BL, SH, AR, GR, TI, SG, NE and JU. For the remaining cantons, Expat's votes were excluded. The original goal to exclude Expat's votes was not achieved, and it is hard to estimate how this could affect the analysis results, as it was not possible to determine how many Expat's were egligible to vote in the '1:12 - for fair wages' Popular Initiative and how many of them actually voted.

Data Flow

Data science project Data Flow

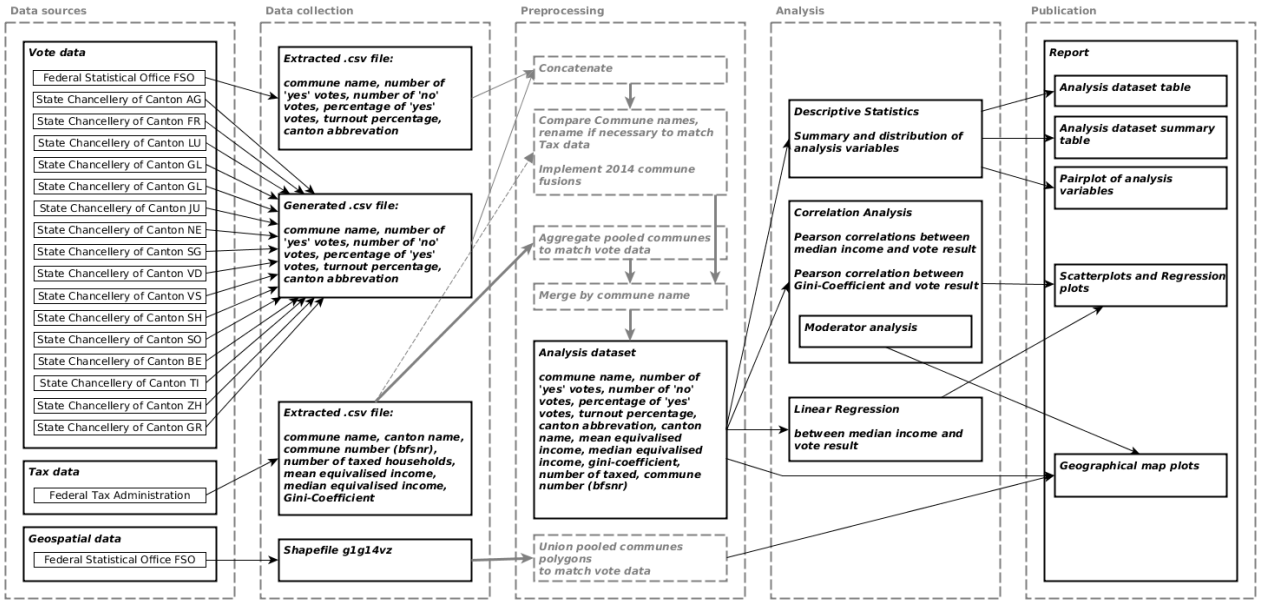


Figure 2: The Data Flow Diagram of this data science project.

The file 'data_sources.txt in the 'Metadata' folder contains all the links to the data sources. As described in Chapter 'Data', data collection involved manually collecting and inserting voting data into one single .csv file - raw data however is provided in the 'Raw_data' folder. Some preprocessing was also done manually (especially renaming communes to match tax data). The analysis process is portrayed in the Jupyter Notebook of this project. All plots and tables can also be found in the 'Publications' folder.

Project workflow

The following Figure 3 portrays the general workflow of this data science project.

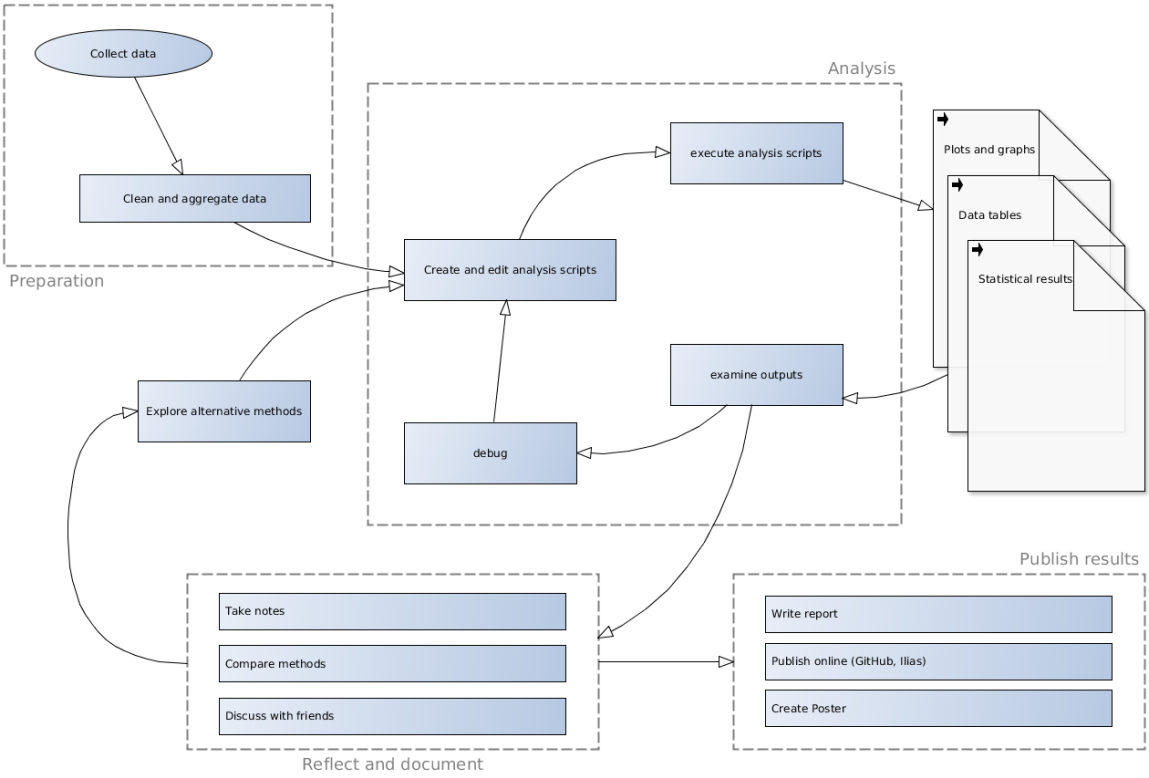


Figure 3: Project workflow of this data science project.

Data Models

Conceptual Data Model

In this project vote and tax data was aggregated into one dataset, namely 'analysis_dataset.csv' containing only one 'entity'[17]. Thus portraying this data as a conceptual data model would not provide any information and it's creation doesn't seem to make sense.

[17] That would be the commune

Logical Data Model

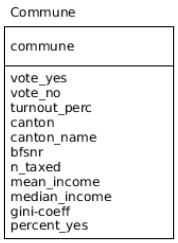


Figure 4: Logical Data Model

Physical Data Model

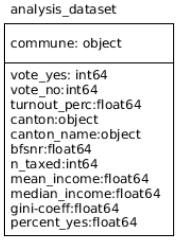


Figure 5: Physical Data Model. The 'analysis_dataset' requires 219.7+ KB of memory.

Risks

The following potential risks were identified:

- **Data loss**: There is a potential risk of data loss, that is the loss of raw data, preprocessed data, analysis scripts, report and poster drafts, output files and plots. To counteract this, Backups of the data were stored on a personal NAS whenever it seemed appropriate.
- **Unauthorized access**: There always is a minimal potential risk of unauthorized data access. To prevent potential attackers from messing up this project, all data was always stored on fully encrypted devices, VPN was used to syncronize Backups on the personal NAS and for the duration of the project only WPA2 secured wifi connections were used.
- **Data collection errors**: There is the risk that when data was collected, mistakes were made in the process of copy-pasting vote data into the LibreOffice Calc csv sheet. By the time this report was written, a more suitable way to collect the data was found, see Chapter 'Data' for further information.
- **Data analysis mistakes**: There is the risk of potential bugs and flaws in the data analysis script, use of methods and of course within the used python libraries and functions.
- **Dataset risks**: As already mentioned in the Chapter 'Data Quality', Expat's vote data could not be excluded from the dataset for all cantons. This could impact the analysis results.

Preliminary Studies

Correlations

The main goal of this data science project was to examine the relationship between the acceptance of the Popular Initiative "1:12 - for fair wages" in Switzerland in the year 2013 and income as well as income inequality on a communal level. Therefore, correlations between these variables are examined. Shapiro tests indicate that the variables used in the correlational analysis are normally distributed[18], therefore the pearson correlation method is applied.

[18] See Pmueller_project.jpynb

	mean_income	median_income	gini-coeff	percent_yes
mean_income	1.000000	0.406749	0.428746	-0.220348
median_income	0.406749	1.000000	-0.077555	-0.389167
gini-coeff	0.428746	-0.077555	1.000000	-0.056398
percent_yes	-0.220348	-0.389167	-0.056398	1.000000

Table 3: Pearson correlations between the analysis variables.

Gini-coefficient and vote result were not correlated (Pearson, r=-.056). Median income and acceptance of the Popular Initiative are slightly negatively correlated with a pearson correlation of r=-.39.

Linear Regression

Simple linear regressions were calculated to predict 'Percent yes' based on median income and Gini-Coefficients. A significant regression equation was found ($F(1,2341) = 417.8$,$p<.001$), with an R^2 of .151. Commune's percent 'yes' decreases by 0.0004 whenever median income is increased by 1 CHF. The relationships between these variables are further illustrated in Figure 6.

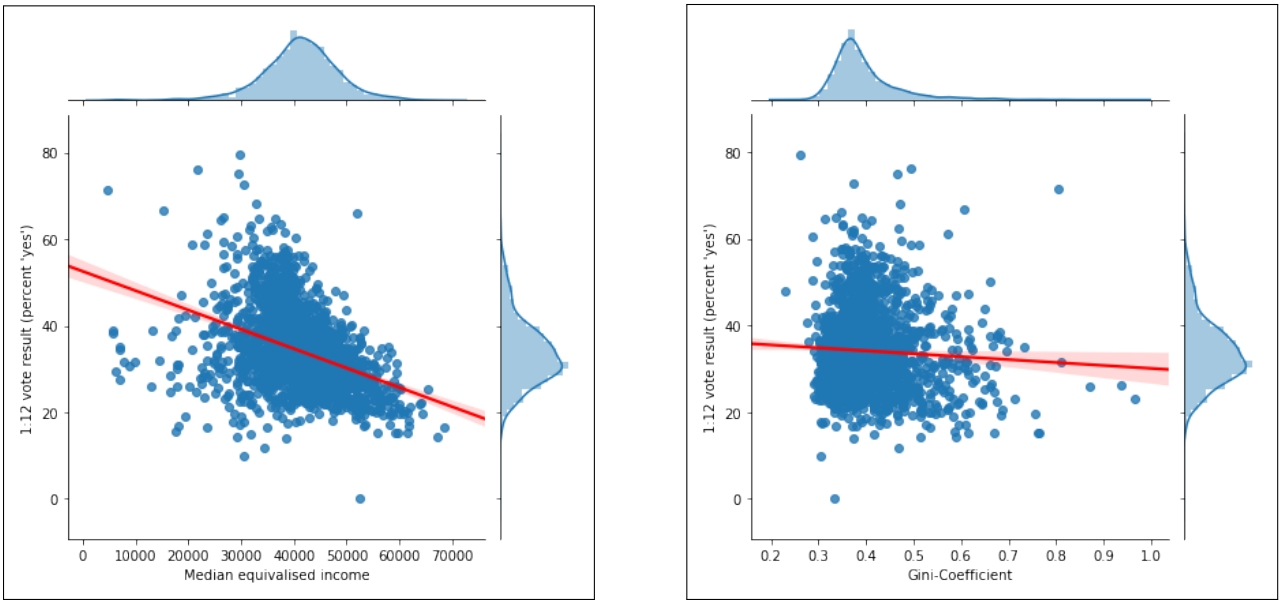


Figure 6: Bivariate Scatter-Plots of median income and vote result respectively of Gini-Coefficient and vote result with fitted linear regression line.

Moderator Analysis

To assess if and how this correlation between mean income and vote result differs between cantons, correlations on canton-level were calculated (see Figure 7). Although the negative correlation between

canton	r =	p	canton	r =	p
BE	-0.31	< 0.001	**UR**	-0.09	0.69
ZH	-0.71	< 0.001	**GL**	-0.51	0.66
GR	0.13	0.11	**AI**	0.76	0.08
BS	-1	0.02	**AR**	-0.05	0.82
BL	-0.65	< 0.001	**LU**	-0.36	< 0.001
TG	-0.35	0	**SG**	-0.54	< 0.001
AG	-0.36	< 0.001	**NW**	-0.61	0.05
VS	-0.1	0.26	**OW**	-0.01	0.98
VD	-0.46	< 0.001	**SZ**	-0.71	< 0.001
NE	-0.29	0.08	**ZG**	-0.26	0.44
FR	-0.3	< 0.001	**TI**	-0.31	< 0.001
SO	-0.26	0.01	**JU**	0.1	0.44
GE	-0.74	< 0.001	**SH**	-0.168	0.412

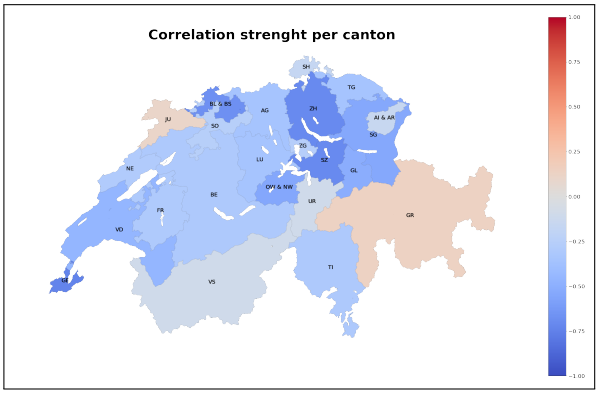


Figure 7: Correlation strenght per Canton.

income and acceptance of the Popular Initiative seems to be an almost nation wide trend, it's strength varies greatly between

cantons. "Canton" could thus be considered a **moderator variable**.

Conclusions

The analysis results indicate that the acceptance of the "1:12" Popular Initiative varies with income but surprisingly not so much with income inequality. Median income per commune seems to be a more reliable indicator of income height per commune than the mean income, most likely because of a small percentage of extremely rich individuals. At least this seems to be the explanation for the extreme values in the commune of Anières in the canton GE, which was in the year 2014 not only the 'richest' but also the most unequal commune in Switzerland[19]. Although the correlation between median income and vote result is slight at most and moderated by canton, it nonetheless indicates an interesting association.

[19] See https://www.srf.ch/news/schweiz/120-millionen-zu-viel-im-steuerkaesseli

But correlation does not equal causation and no explanation for this relationship can be provided at this point, thus further analyses are needed to investigate the relationship between income and voting behaviour. Other variables that were not included in this analysis should also be taken into consideration, such as political party strength or people's orientation towards political convictions

References

Data Sources

- Federal Statistical Office FSO
- Federal Tax Administration
- State Chancelleries of Cantons AG, FR, LU, GL, JU, NE, SG, SH, SO, TI, VD, VS, GR, BE and ZH