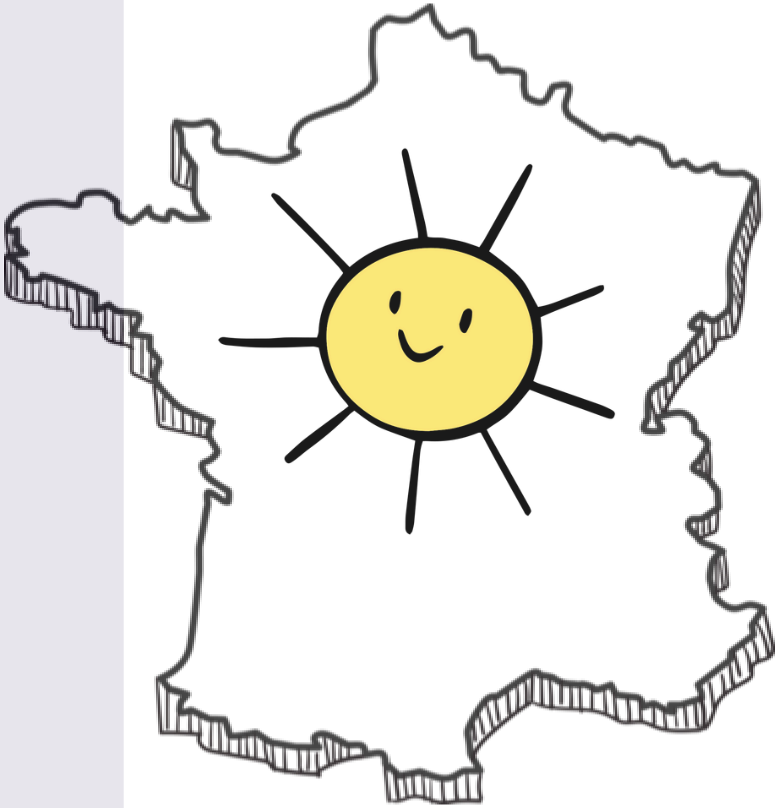
FRENCH TAXES

~ and other fun topics ~

Calvin Davis
https://github.com/MyNameIsCalvinDavis/



Index

Introduction

In this report, we aim to evaluate the French fiscal tax dataset provided through Desights.ai in order to understand and answer the following questions:

- Context for French Taxes & Our Dataset
    - What data does our dataset contain? What does it mean?
    - What is worth evaluating in our dataset? What do we consider important features?
    - Of the important features we identify, what do they control, and how much do they contribute to a commune's income?
    - What laws, policy, or other guidance can we find to support our conclusions?
    - What assumptions or conclusions can we make about the quality of our dataset?
- Municipality Rankings & Revenue Trends
    - What are the top and bottom ranked communes by income in this dataset?
    - How do these communes compare to each other? How do they change over time?
    - How does revenue change over time? Are the top earners always in the top?
        - Visualize the past 5, 10, 15, and 20 years of income data for top earning communes, and all communes
    - How does commune income grow or shrink over time?
    - How did the Professional Tax Reform change French GDP / commune income after 2010?
- Data Analysis
    - How closely related are population and total income for a province?
    - Are there other variables that predict province income better than population?
    - Can we create a machine learning model to predict future income? What are the pros/cons of such a model?
    - What taxes contribute the most to commune income? Do they apply to all communes?

Dataset Description

The ***French Fiscal AI Innovation and Prediction Challenge***, held in collaboration with the **Joint Research Centre (JRC),** the Science Hub of the European Union Commission, presents a unique opportunity to examine **40 years of French tax data**. The JRC, known for its commitment to providing rigorous, evidence-based scientific and technical support to EU policymakers, is partnering to enhance data-driven decision-making. Participants will employ their analytical prowess and predictive modeling skills to reveal significant insights into the fiscal behaviors of French municipalities.

The challenge invites an in-depth exploration of the data, beginning with ranking municipalities by their tax revenue to identify those with the highest income from direct local taxes. Participants will create graphical representations to illustrate revenue trends in selected towns, drawing meaningful conclusions from these visualizations. The analysis extends to classifying municipalities based on their growth in tax revenues over various periods, investigating the relationship between population size and tax revenue, and assessing the impact of eliminating the professional tax collected before 2010. This comprehensive approach aims to uncover significant insights into the fiscal dynamics of French municipalities.

Understanding French Tax Law

Understanding the context of our data is where a majority of the work of this challenge comes from. We are provided with a spreadsheet that contains information about the columns and what they represent according to the data source, translated by ChatGPT into english (*data_descriptions.csv)*. One issue is that besides a small handful of intuitive column names (region / commune / dept names, population, etc) a lot of the data is unintuitive.

Goals:

1. Identify relevant variables in our data

…and that's it, really.

Identifying Relevant Variables

This handy chart I found online outlines a general overview of the French Tax System, and breaks down specific categories quite well. Good bedtime reading material.

https://www.impots.gouv.fr/sites/default/files/media/3_Documentation/brochures/french_tax_law_brochure_2024.pdf

Inside it breaks taxes into a few categories:

- Business Profits
- Non commercial profits (not trader status)
- Agricultural profits
- Property profits
- Wage, salaries, pensions, annuities
- Investment
- Capital gains

Based on the descriptions of the cells in our data explanation spreadsheet, it seems we're working specifically with the Local Direct Tax System, which describes four taxes:

- Property Tax on Developed Land (TFPB)
- On undeveloped land (TFPNB) (TAFNB)
- Residence Tax (TH)
- Local Economic contribution (CET)
    - Business premises contribution (CFE)
    - Business value added (CVAE, wont exist in 2024)

Despite the French government's suggestion that there are four base taxes, there are actually a great deal many more. Please understand: I am not a French Tax consultant. What I'm saying may be obvious.

Reading the scary spreadsheet (*data_descriptions.xlsx)*

Many variables seem to be clumped together to represent the same specific tax, breaking down specific exemptions for different groups, regions, departments, applicable laws, etc. I want to assume that these variables are included for the sake of completeness.

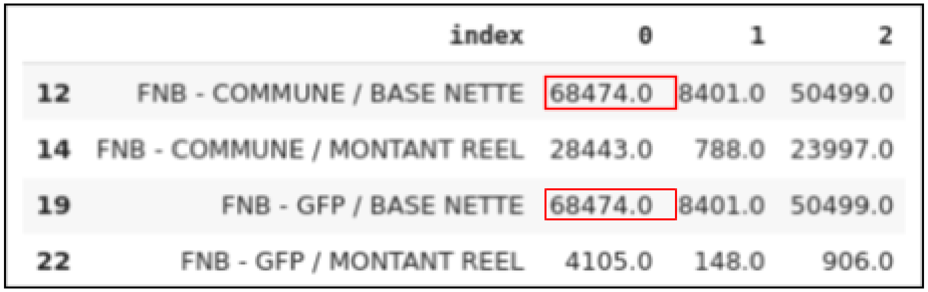
The parts we care about are a tax's specific **base value** and **actual / real value**. Breakdowns of the base tax, which comprises most of the data, include specific exemptions based on local municipality laws, regional considerations, revenue streams for specific taxes and where they go and who they go to, EPCI regional decisions, individuals within certain tax brackets, whether the land is subject to things like GEMAPI (a land-water tax), etc. Basically, a bunch of stuff we don't care about. **Importantly, the base tax includes these exemptions** already, so we don't need to worry about factoring them in ourselves.

After looking through this dataset for some time, I have observed the following:

- Each tax category has a mathematical tax base and an actual taxable amount, usually with a large discrepancy
    - **This discrepancy is often due to "gains / losses"** or other specific tax law that I'm sure isn't reflected in this dataset
    - **Each tax is further broken down** into rate determinations, "vote rate", # of taxable items (in the region?), rates for this tax based on income or other laws, adjusted values based on specific regional context, # of exempted individuals / houses / businesses, etc.
    - **Smoothing rate exists, which we will be ignoring**: *The aim of smoothing is to gradually integrate (over 10 years) the effects of the reform of rental values for business premises. A smoothing amount was calculated in 2017 for all business premises existing on January 1, 2017, and is applied (upwards or downwards) to the TFB property tax assessment until 2026.* **We will ignore this rate because its final value is reflected in the "actual" amount anyway.**
- **Some of this data is unrelated to tax** and describes the number of facilities that fall into certain categories (like 535 - Anti pollution facilities / number of items). There is also a great deal of categorical data, like whether a commune falls into a specific environmental category / department / is eligible for some exemption, etc.
- A lot of this data describes quantities of exemptions (presumably, lost tax from the base rate)
    - For example: Without a further understanding of CFE and a tax advisor it's not realistic to subtract (EXEMPTION OF S.A.I.C. (col 558)) from the total CFE tax levied against a region.
    - Col 12 (TFNB tax base) describes: **"The tax base is a net base that takes into account any applicable allowances and exemptions."**. This is true for the rest of the "Tax Base" columns as well.

As such, our assumptions are as follows:

- The only applicable fields for broad-stroke analysis of regions are tallied in Base Tax / Actual Amount fields
- Most of this data isn't relevant unless we want to analyze specific breakdown criteria for a tax category accross France



	index	0	1	2
12	FNB - COMMUNE / BASE NETTE	68474.0	8401.0	50499.0
14	FNB - COMMUNE / MONTANT REEL	28443.0	788.0	23997.0
19	FNB - GFP / BASE NETTE	68474.0	8401.0	50499.0
22	FNB - GFP / MONTANT REEL	4105.0	148.0	906.0

For example, here we have three random communes (each column a different commune) and two sub-categories of FNB, describing TFNB, a property tax. With TFNB tax, the tax base (68474.0) is actually described in two columns (with the same value), one of which seems to be the federal rate, and the other a local rate (EPCI). We see here a breakdown of this tax:

Col14 describes its data as such:

> TFNB issued to the municipality
>
> The actual amount corresponds to the sum of TFNB contributions due by taxpayers on the **municipality's** territory. The difference between the actual amount and the mathematical product of base (B11) x rate (B12) is due to gains and losses.
>
> The tax paid by each taxpayer is rounded off to the nearest euro. For this reason, the sum of taxes payable by taxpayers may differ slightly from the simple application of the voted rate to the municipal base (revenue paid to the local authority). If the actual amount is higher than the mathematical product, this will be a gain for the State and therefore a loss for the local authority, and vice versa.

Col 22:

> TFNB issued to the EPCI with its own tax authority on the municipality's territory.
>
> The actual amount corresponds to the sum of TFNB contributions due by taxpayers on the **commune's** territory. The difference between the actual amount and the mathematical product of base (B11) x rate (B12) is due to gains and losses.
>
> The tax paid by each taxpayer is rounded off to the nearest euro. For this reason, the sum of taxes payable by taxpayers may differ slightly from the simple application of the voted rate to the municipal base (revenue paid to the local authority). If the actual amount is higher than the mathematical product, this will be a gain for the State and therefore a loss for the local authority, and vice versa.

This provides additional context and reinforces the idea that the sum of "actual rates" over the dataset is probably the way to go. In this example, **it wouldn't make sense to sum both base rates** since they pull from the same source, but the description suggests that it's okay to add both "actual amounts" because EPCI has its own tax authority.

It's worth noting that the descriptions of the "actual amount" columns are inconsistent. Some (14, 22) as:

> The actual amount corresponds to the sum of [...] contributions due by taxpayers on the municipality's territory. The difference between the actual amount and the mathematical product of base [...] is due to gains and losses.

And some (cols 18, 34, 46) as simply:

> Proceeds from [...] issued to the syndicate on the commune's territory.

I don't think this changes anything, as both describe a concluded sum.

Non-useful Data

A lot of this data seems to just be informative, and not used in calculations. For cols 1149,1150, describing income brackets:

Minimum CFE base for taxpayers with sales of €10,000 or less.

Presumably there's no way that we could calculate the total revenue from this group (and others) of taxpayers, unless we also had a number of people in this tax range. A lot of this data *seems* to exist (maybe) in cells 741-794, though unfortunately 738-882 has no description of this data. And even if it did, the math probably isn't as simple as taking the *(number of people in a category) * (the tax per individual in that category)*. We can only **assume such granularity is either already factored into the base values or actual values** of each tax, mentioned earlier.

Sometimes, like in col 42, we get descriptions like this:

42 *Proceeds from GEMAPI backed by TFNB accruing to EPCI with its own tax authority on the commune's territory*

To me this means, money comes out of the TFNB bucket, renamed / redirected to GEMAPI classification, and then sent to EPCI. This raises questions:

- Is the real amount of TFNB revenue modified by this reduction? Is it considered a reduction at all, or does GEMAPI fall under TFNB?
- GEMAPI (a water land tax) probably does not fall under TFNB (an undeveloped land tax) but technically the tax for undeveloped land is called either TFPNB or TAFNB, so maybe it's different?
- Why does GEMAPI not have its own classification? Should we group its actual amount under TFNB?

Conclusion

As we will discuss and show later, little issues like this don't actually matter that much in our dataset, even though there will be a lot of them. The reason for this is that we are evaluating total income of a commune in various forms, and the impact of a single tax, even if large, will affect everyone equally for large taxes. Specifically, something like a business levy tax will hit communes proportionally, and so shouldn't change any of our conclusions.

This might be a bigger issue if we were breaking down the specific exemptions, rates, and funds of each tax, which we will not do here.

Selected Columns & Descriptions

The data we decide to use for our dataset is, unfortunately, something I have manually combed through and read descriptions for in order to identify relevance. Based on our findings above, we decide to move forward by capturing all of the "actual value" data, and ignoring everything else. If something seems relevant, I include it (or at least mention it if it is not included in our selection). **Most of these are the "Real Product" or "Actual Amount" of the tax category they represent.**

10	**Name**
1180	**INSEE CODE**
1181	**YEAR**
1120	**REGION CODE**
1145	**DEPARTMENT NAME**
1146	**REGION NAME**
1147	**SIREN NUMBER**
1148	**Name of EPT to which commune belongs**
1090	**Population**
- -	
14	FNB - COMMUNE / MONTANT REEL
18	FNB - SYNDICATS ET ORG.ASSIMILES / MONTANT REEL
22	FNB - GFP / MONTANT REEL
25	TAFNB - COMMUNE / MONTANT REEL NET
34	FNB - TSE / MONTANT REEL
38	FNB - TSE GRAND PARIS OU EPFL GUADELOUPE OU EPFL MARTINIQUE / MONTANT REEL
42	FNB - GEMAPI / MONTANT REEL INTERCOMMUNALITE
46	FNB - CHAMBRE D'AGRICULTURE / MONTANT REEL
53	FNB - CAAA / DROIT PROPORTIONNEL - MONTANT REEL
69	FB - COMMUNE / MONTANT REEL
74	FB - SYNDICATS ET ORG. ASSIMILES / MONTANT REEL
79	FB - GFP / MONTANT REEL
83	FB - TSE / MONTANT REEL
88	FB - TSE AUTRES / MONTANT REEL NET
93	FB - GEMAPI / MONTANT REEL INTERCOMMUNALITE
98	FB - TASA / MONTANT REEL
105	FB - TAXE D'ENLEVEMENT O.M. / TAUX PLEIN - MONTANT NET LISSE
●	**TEOM revenue at full rate - there are other rates we will be ignoring due to lack of information**
122	FB - TAXE INCITATIVE ENLEVEMENT DES ORDURES MENAGERES / MONTANT REEL / COMMUNE
123	FB - TAXE INCITATIVE ENLEVEMENT DES ORDURES MENAGERES / MONTANT REEL / SYNDICAT
124	FB - TAXE INCITATIVE ENLEVEMENT DES ORDURES MENAGERES / MONTANT REEL / GFP
176	TH - COMMUNE / MONTANT REEL DONT THP/E AU PROFIT DE L ETAT
177	TH - COMMUNE / MONTANT REEL COMMUNAL DE THP/E AU PROFIT DE L ETAT

178	TH - COMMUNE / MONTANT REEL COMMUNAL AU PROFIT DE LA COMMUNE
186	TH - SYNDICATS ET ORG. ASSIMILES / MONTANT REEL
193	TH - INTERCOMMUNALITE / MONTANT REEL DONT THP/E AU PROFIT DE L ETAT
194	TH - INTERCOMMUNALITE / MONTANT REEL INTERCOMMUNALITE DE THP/E AU PROFIT DE L ETAT
195	TH - INTERCOMMUNALITE / MONTANT REEL INTERCOMMUNALITE AU PROFIT DU GROUPEMENT
202	TH - TSE / MONTANT REEL
207	TH - TSE GRAND PARIS OU EPFL GUADELOUPE OU EPFL MARTINIQUE / MONTANT REEL
212	TH - MONTANT REEL INTERCOMMUNAL TAXE GEMAPI
333	CFE - COMMUNE / PRODUIT REEL NET
347	CFE - INTERCOMMUNALITE / PRODUIT REEL NET / FP UNIQUE OU EN ZAE
352	CFE - INTERCOMMUNALITE / PRODUIT REEL NET / FP UNIQUE
357	CFE - INTERCOMMUNALITE / PRODUIT REEL NET / FPZ EN ZAE
362	CFE - INTERCOMMUNALITE / PRODUIT REEL NET / FPE EN ZONE EOLIENNE
368	CFE - TSE / PRODUIT REEL NET
372	CFE - GEMAPI / PRODUIT REEL NET INTERCOMMUNAL / FISCALITE ADDITIONNELLE
376	CFE - GEMAPI / PRODUIT REEL NET INTERCOMMUNAL / FP UNIQUE OU EN ZAE
380	CFE - GEMAPI / PRODUIT REEL NET INTERCOMMUNAL / FP UNIQUE
384	CFE - GEMAPI / PRODUIT REEL NET INTERCOMMUNAL / FPZ EN ZAE
388	CFE - GEMAPI / PRODUIT REEL NET INTERCOMMUNAL / FPE EN ZONE EOLIENNE
392	CFE - TASA / PRODUIT REEL NET
396	CFE - CHAMBRE DE COMMERCE ET INDUSTRIE / PRODUIT REEL NET
403	CFE - CHAMBRE DES METIERS / DROIT ADDITIONNEL / PRODUIT NET
869	IFER TOTALE / COMMUNE
870	IFER TOTALE / INTERCOMMUNALITE

- **Flat-rate network company tax (**https://entreprendre.service-public.fr/vosdroits/R14668?lang=en**)**

932	DCRTP / Commune
983	DCRTP / INTERCOMMUNALITE
1044	DCRTP / DEPARTEMENT
1066	DCRTP / REGION
1136	TP - SOMME DES ALLOCATIONS COMPENSATRICES / DEP
1140	TP - SOMME DES ALLOCATIONS COMPENSATRICES / REG

Data Pre Processing, Loading, Cleaning

This data comes from a **3.6GB** CSV file, pared down to a **0.6GB** parquet file, a common storage format for columnal data and relational databases.

Steps:

1. Load our data with the provided DuckDB API and download a local working copy of the parquet file
2. Identify what columns exist in the database and get a sample of them to work with
    a. Realize there are too many columns to fit in the dataset in memory and we'll need to be more creative, as my machine had 8GB ram only
    b. See Analyzing French Tax Law
3. Begin gathering necessary information about specific communes

Most of this process is in code format - refer to the python notebook in the appendix if you'd like a more granular understanding of how this was done.

Cleaning

We've chosen not to clean this dataset for several reasons:

- Very few of the rows (communes) or columns (specific taxes) have no missing data. Between missing values, 0-filled data, infinite values, etc., there is too much imperfect data to clean without destroying a significant portion of the dataset.
- Real world data isn't any less valuable just because it doesn't have all 40 years of information associated with it
- Some taxes aren't applicable to certain communes, so the data will reflect 0 or NaN, which is accurate

In any event, this shouldn't really matter - most of our analysis will involve grouping, which alleviates the issues of data accuracy / consistency a little bit by taking a look at overarching trends instead of specific values.

Validation

Were the columns we selected accurate? Let's find out. To start, we take a random row, and see what we can find out about its data. In our example, we choose a row at random, Ambronay 2022.

A quick look at this row in our data shows the following:

```
AMBRONAY 2022
Sum of taxes: 3900874.0 --- Population: 2915.0
1338.20 per individual
```

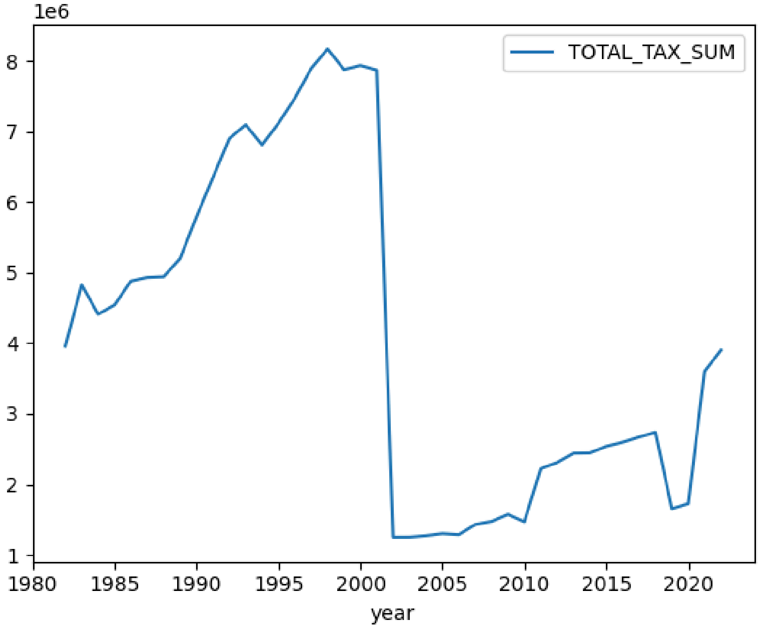
Some wikipedia information for this commune (https://en.wikipedia.org/wiki/Ambronay#Economy) for 2008:

> **Elements of calculation**
>
> - Total population DGF = **2,316**
> - Potential 4 taxes = 1,250,326
> - Financial potential = **1,533,145**
> - Financial potential by population DGF = ~ 661.98
> - Financial potential per capita stratum = ~ 727.92
>
> **Main financial resources**
>
> - Buildings (before capping) = 255,016
> - Undeveloped land (before capping) = 30,713
> - Housing tax (before capping) = 200,943
> - Block grant = 499,653
> - assets of Urban solidarity (DSU) = 6,776
> - Total assets (Lump - DSU - DSR - DNP) = 506,429
> - Total per population by DGF = ~ 218.66
>
> The working population is about **1,000** with an unemployment rate above **9%**. The commune has 82 business enterprises with core businesses being trades, construction, and particularly services.

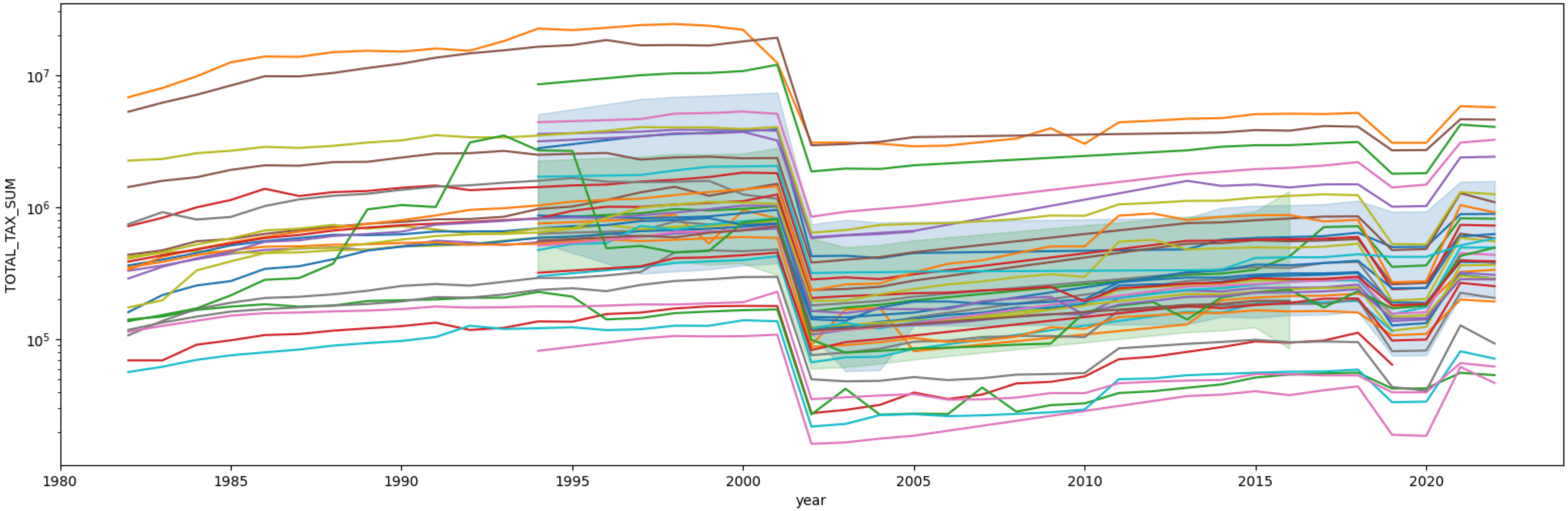
Looks like our selection roughly aligns with this data. From 2008 to 2022, there appears to have been a **26% increase in population** (2316 > 2915) and approximately a **140% increase in revenue** (1.533.145 > 3.900.874). This implies a significant development over these 15 years or so. This is good for us because it means our data is, at the very least, within the same ballpark. We could also interpret this to mean our margin of error is around 140%, although several factors reduce this error: inflation and a rise of about ~3mil people in French Population from 2008 > 2022, not counting any growth this area has seen in tourism / wealth / development / businesses, etc, allows us to say our data looks good and our selected columns approximate the income for a region accurately.

Big Drop - Francs and Euros

Let's graph Ambronay's tax sum over time:



We see a pretty big drop. We actually see this drop for every value in our dataset around 2001:



There are three interpretations for this issue:

The tax values we selected originally were incomplete and do not accurately represent the data

- We can surmise that there's a missing tax somewhere that we forgot to include, which changed between 2001 and 2002. Perhaps the tax revenues from our other variables were funneled into this missing tax in 2002, or distributed in some other way we did not capture.

The data simply does not explain the dip in tax revenue

- This isn't a bad thing for multiple reasons. Primarily because we can still accurately measure municipality growth over time (with a hiccup in 2002), top and bottom earners, and other similar data. The dip seems to have affected everyone equally, and even if it didn't, there isn't anything we can do about it. The only issue we'll have is when feeding this into an ML model later - because ML tends to be of a predictive nature and requires continuous data, we'll use the last half (> 2002) of the data to train it, and ignore the first half. This isn't technically necessary, but will improve accuracy.

Something happened to the french fiscal / tax system in this time period that explains this drop

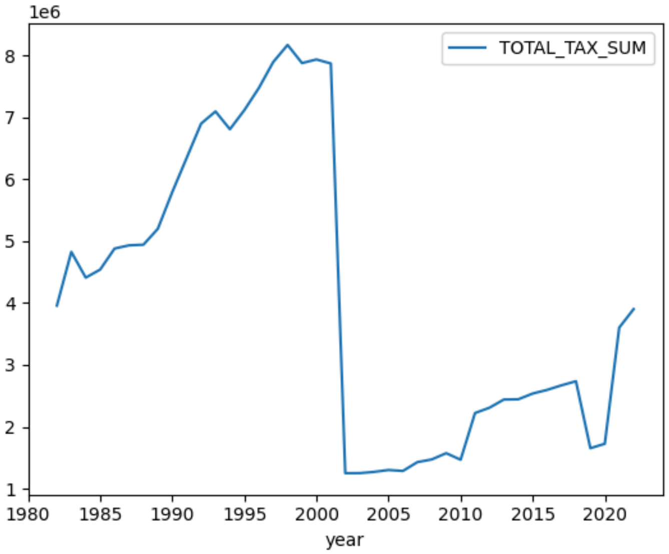
- One hypothesis is that this drop is a conversion rate issue. According to this page (economy-finance.ec.europa.eu), France adopted the Euro 1 Jan 2002, with a conversion rate of **€1 = 6.55957 FRF**.

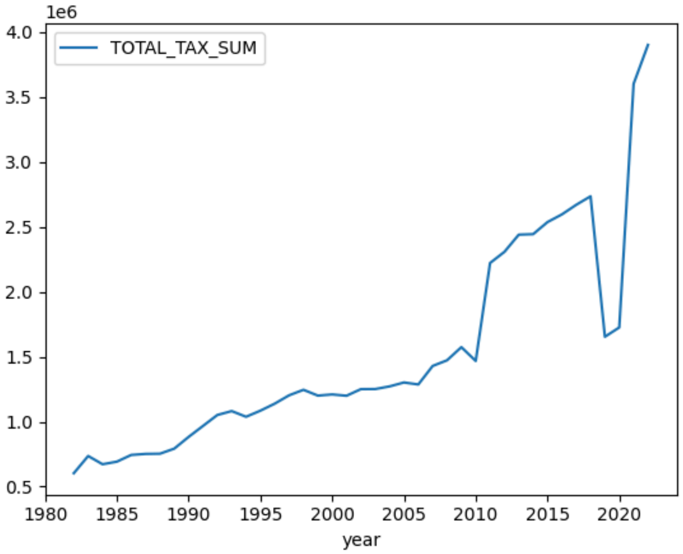
$$Euro\ :\ Franc\ Ratio$$
$$1 = 6.55957FRF$$

$$Ambronay\ Euro\ :\ Ambronay\ Franc\ Ratio\ 2001\ >\ 2002:$$
$$4926508/777836 = 6.334$$

This is almost certainly a conversion rate issue between Francs and Euros. If we use the data from Ambronay above, we can see that the ratios for conversion between the franc and the decrease in income approximately match. So more than likely this large drop in data is probably caused by a currency issue. A **15.2%** ratio and a **15.8%** ratio between these two conversions are close enough, especially considering there was some income change in the province over that single year anyway. A small error is expected.

Indeed - the error seems to be fixed when we apply this conversion rate.





Additional Thoughts

INSEE Codes

Some of the INSEE codes are missing, or change over time. These codes change inconsistently based on local or federal laws, introduction of new communes or elimination of old ones, fiscal systems changing, and other reasons. As such they may be considered one unique identifier of a region, but instead we will just be going with names of communes, since any duplicated tax information between them will be in name only, and in theory communes of the same name will have different tax amounts, so there shouldn't be any problems.

Municipality names

Some of the names are missing, 34 of the names are **#NOM?** which is just missing data. Most of the rest of the names are populated, although there are duplicates, because some regions share a name. This shouldn't be a problem, because in this dataset we are really only exploring the top and bottom of certain statistics (revenue, growth, etc) none of which share names.

Importance of Margin of Error in this dataset

This dataset uses tax information for provinces that derive their income from local laws, federal and local offices, historical records, changes in populations and coding, certain tax breaks, union voting, etc. We assume in this dataset that such inconsistencies between data are both infeasible to remediate at scale, and ultimately not important; for the most part we are identifying trends in the data and applying errors, where they exist, to the entirety of the dataset. As such, if we select a wrong tax, or choose to apply a multiplicative value (like a conversion rate between Francs and Euros, for example), our outcome does not change. If we are choosing a tax to apply, we apply it to the whole dataset.

This may push some results slightly higher or lower on certain graphs (but probably not because each tax is relatively small compared to the total income of a region) which is also not important because our goal is to provide information for the top 5 and bottom 5 communes for various statistics, and in this report we often go up to **10 each** instead of 5. Hypothetically if 5th place got bumped down to 6th because of an oversight or misunderstanding of the data, that commune is still represented / visualized because of our large allowance for such errors in our graphing decisions.

Key Findings

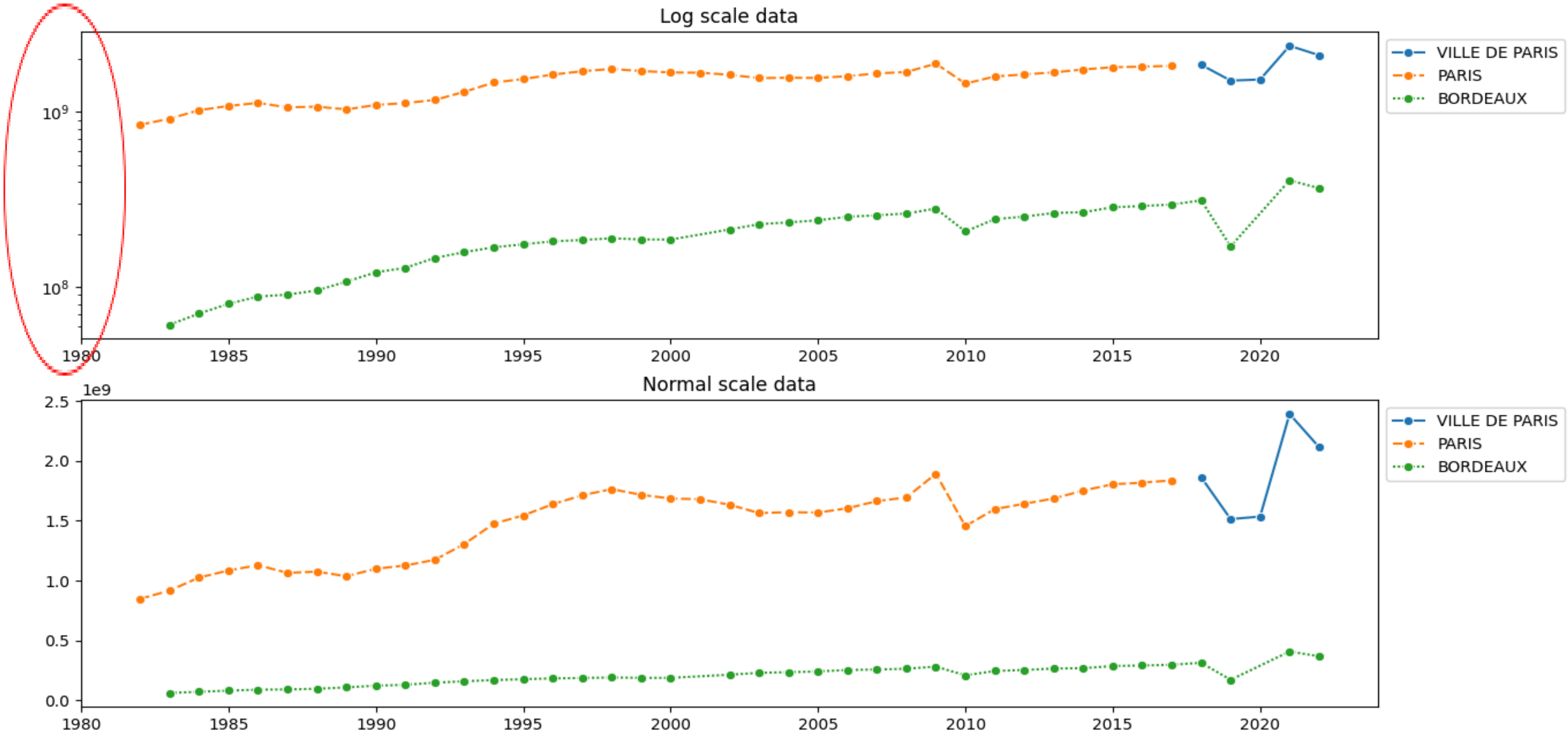
Foreword

As we describe in the section Understanding French Tax Law, a majority of this report's analysis comes from reading and understanding the nuances and context of the dataset itself. Because our data only concerns commune income (and not GDP / debt / retired population / tourism / war / specific budget breakdowns / trade agreements / nontaxable revenue / non-liquid assets like infrastructure / etc), many of the conclusions we draw are very simple, since we really only have one variable to work with (Total Income).

As such, the granular analysis of *how* we reached a specific conclusion or made a specific graph is often omitted, because the answer is either obvious in the graph (the biggest bar with the highest name has the biggest income) or a more specific answer would require a code analysis, which is available in the appendix.

France has many communes - in this report we will often only show the top / bottom 10 of a commune's income, growth, etc. If you would like a more detailed summary or want information about a specific commune, please see the appendix.

Understanding Log-Scale Data



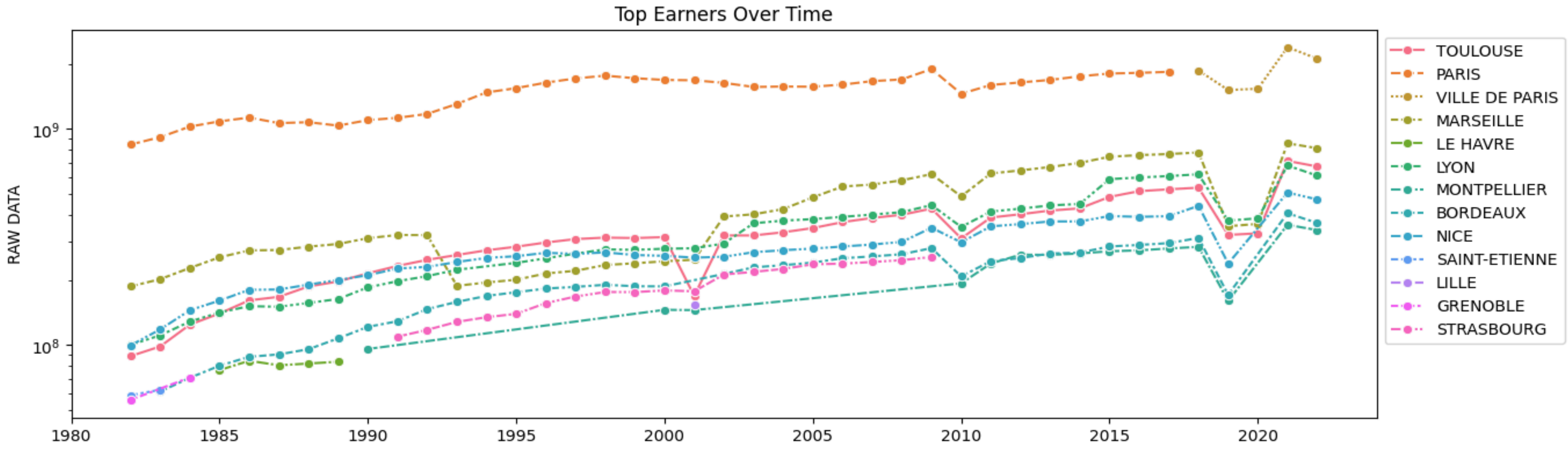
Many of our graphs are in **Log Scale**. This means that data which is very different (**€100 vs €100.000**) can appear on the graph at the same time, and provide similar information. In this example, Paris has an income around **€2.000.000.000** and Bordeaux has an income around **10x less**. The bottom graph demonstrates this on a normal scale - Bordeaux looks flat and doesn't provide very much information. In the top graph, where log scale is used, **we can still find patterns and other dips in our data, despite their difference in scale**.

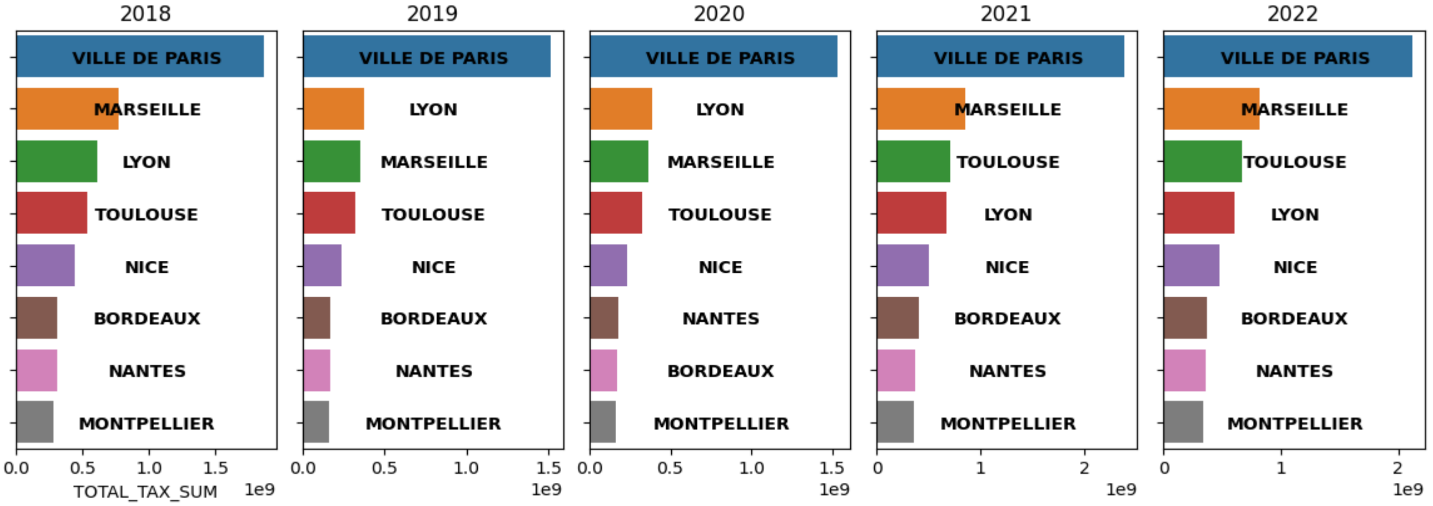
Municipality Rankings - Top / Bottom by Income & Revenue Trends

Location: RAW_DATASET_name_totalsum.csv

For this finding, we sum all of the taxes for each commune, for each year, and represent this value as "TOTAL_TAX_SUM". Graphed, we see how this value changes over time for specific communes.

Paris, **Marseille**, **Lyon**, **Toulouse**, and **Nice** trade places back and forth over time as the top grossing communes. Note: Paris is reclassified to Ville de Paris, and there are some other similar renaming conventions with other, less important communes. For a deeper breakdown of this data, please refer to the csv file mentioned above, which contains specific income information for every commune in France over the last 40 years.

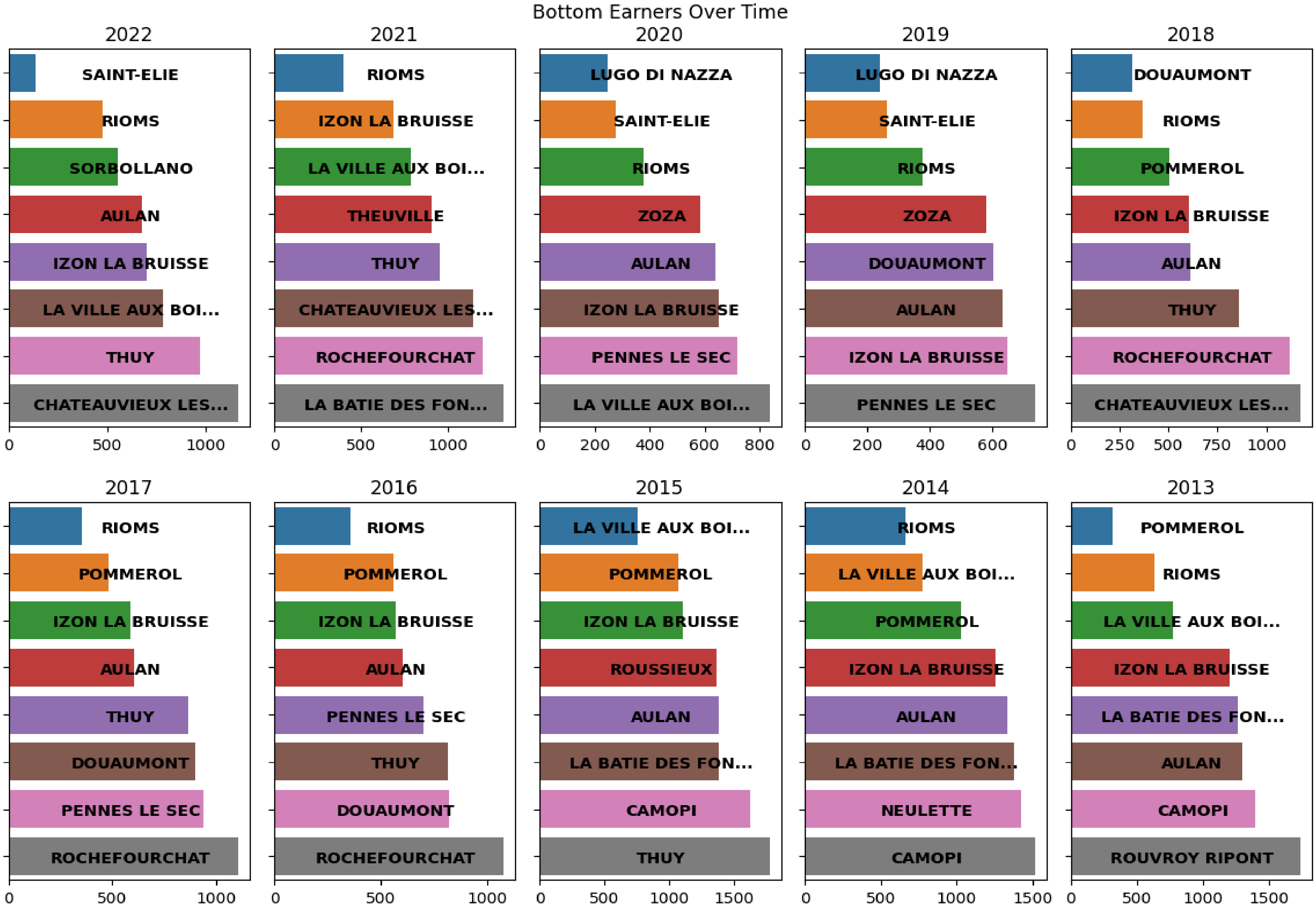




Municipality Rankings - Top / Bottom by Income & Revenue Trends

Location: RAW_DATASET_name_totalsum.csv

Bottom earners change so frequently that it's not worthwhile making a time series graph. Instead we just provide a Top 8 breakdown for the last 10 years. Our data changes over time, so this visualization is the best we can do without more specific evaluation criteria. The csv mentioned above contains specific income information for every commune in France over the last 40 years.

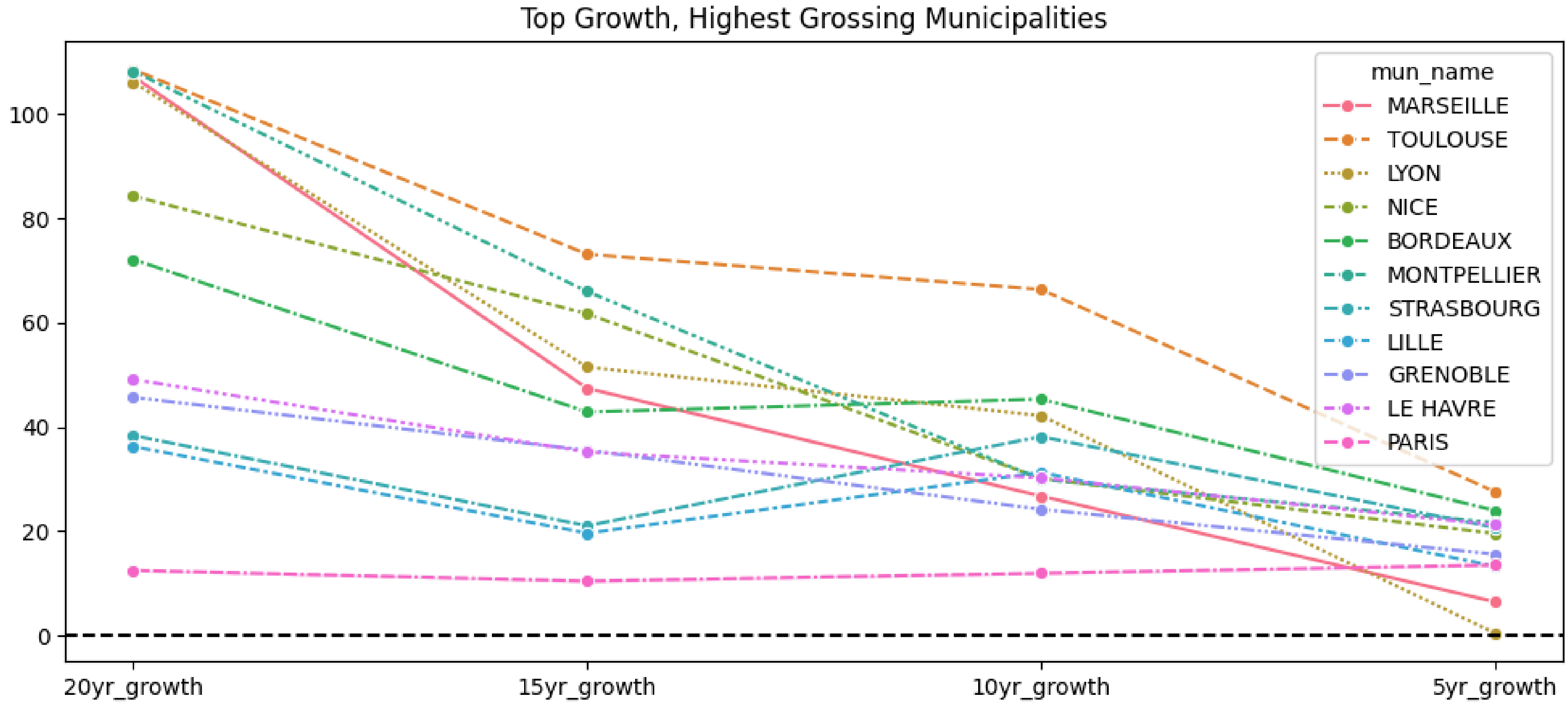


Municipality Rankings - Top / Bottom by Growth

Location: total_growth.csv

To calculate growth, we take the category we want to analyze (5yr_growth, for example) and find 2022 income, and 2022 - 5 income (2017). We take the two incomes and calculate a percent increase, and then record that data for a specific commune. Each commune does this four times, for 5,10,15,20 years of data. This means that **growth is not from one category to another** (15yr > 10yr, as the graph would imply) but rather a comparison to 2022's data.

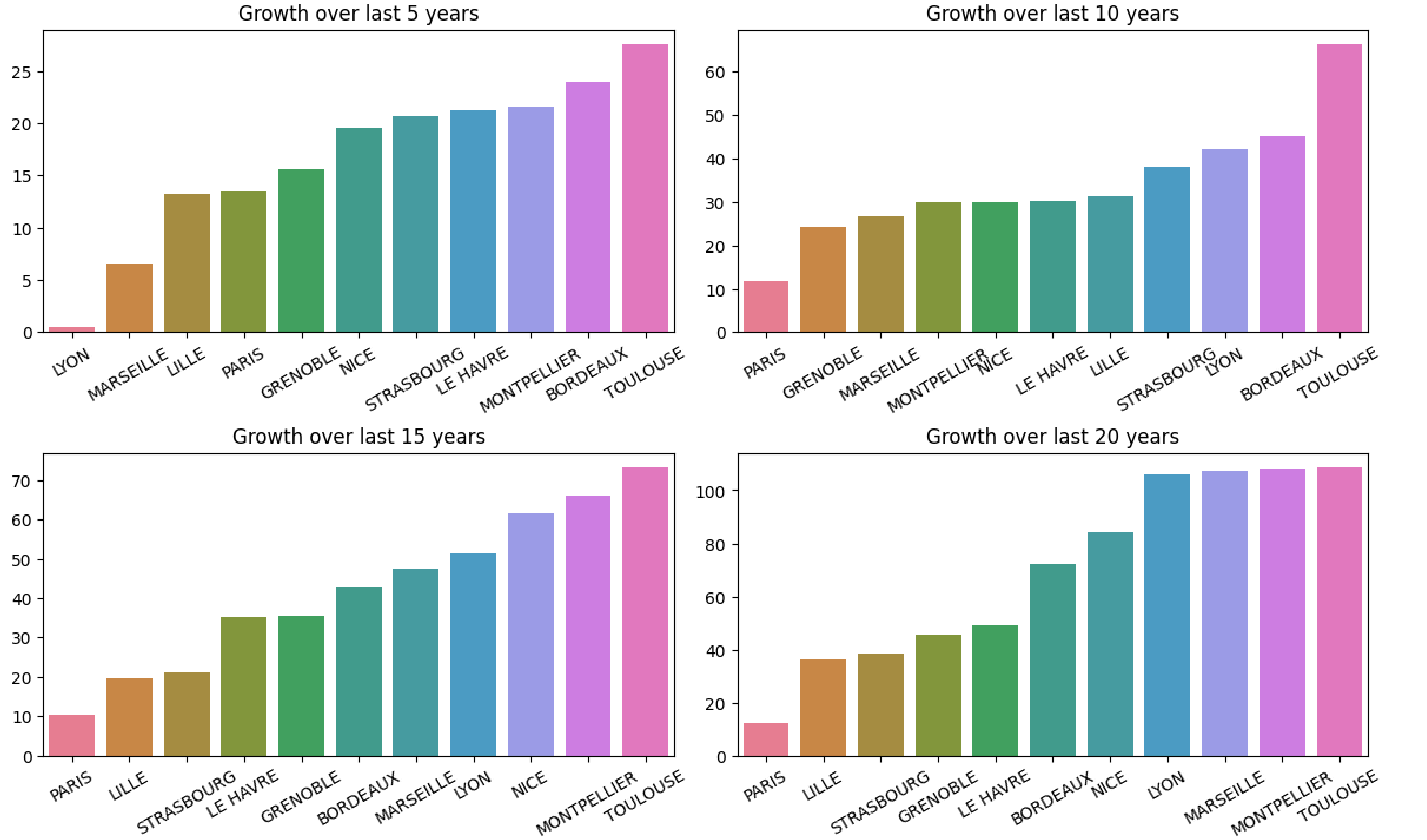
Here, we breakdown growth specifically for the top communes identified above. These are not necessarily the top growing communes in all of France. All of the top grossing communes have seen significant growth in the last 20 years, except **Lyon** which saw neutral growth in the last 5 years.



Municipality Rankings - Top / Bottom by Growth

Location: total_growth.csv

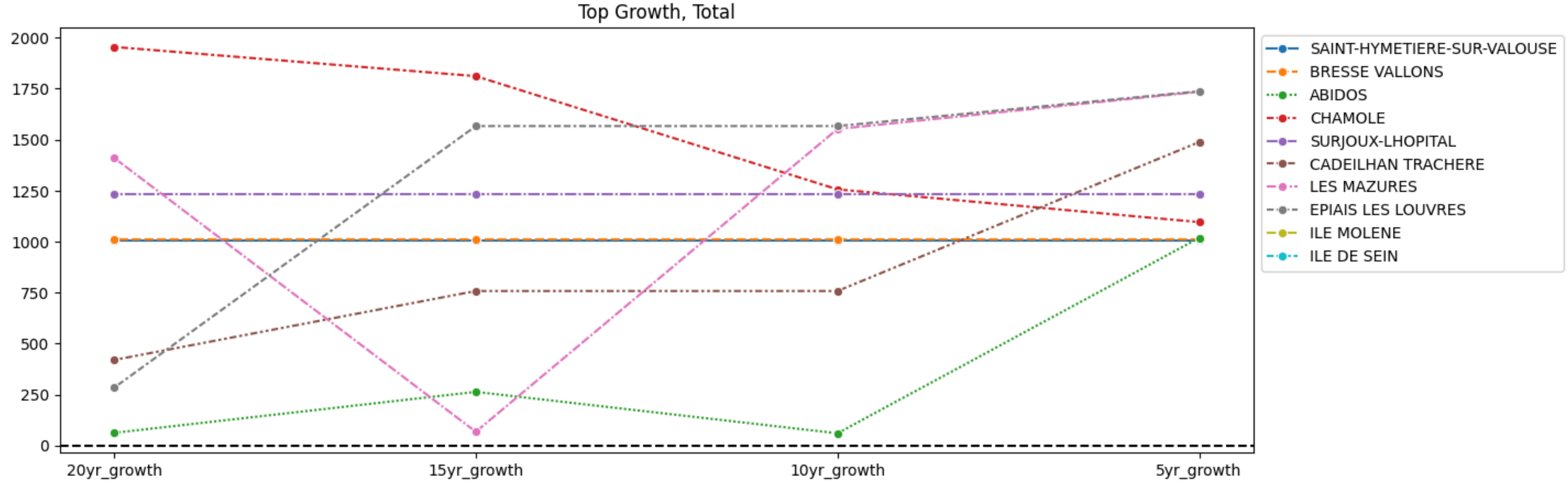
The same visual provided above, in bar graph form. **Note**: None of the top grossing communes see negative growth in the last 20 years.

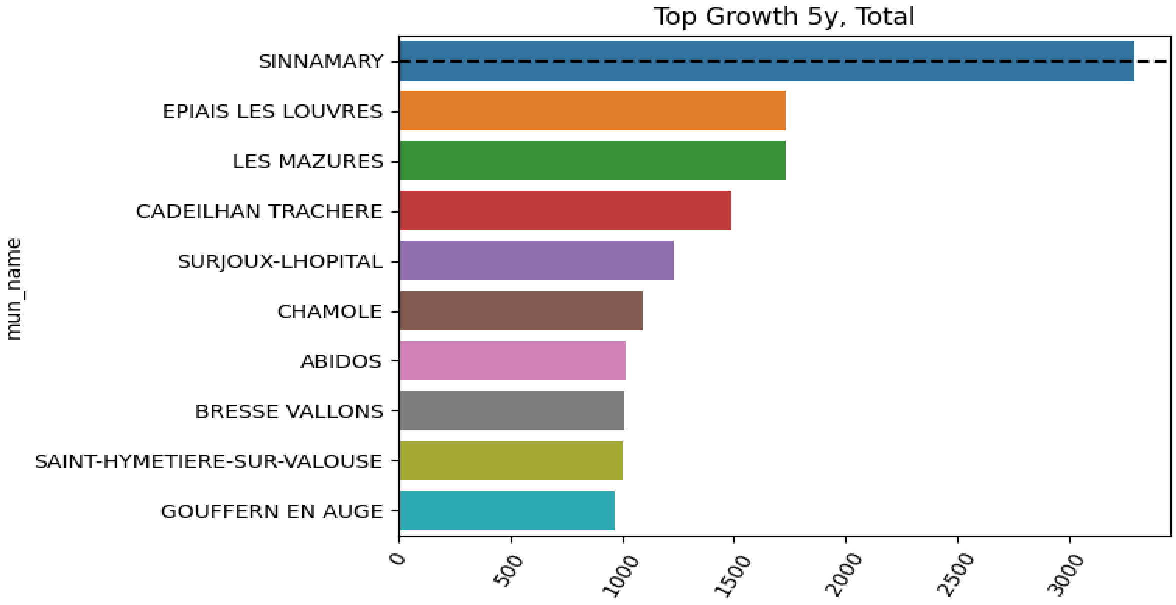


Municipality Rankings - Top / Bottom by Growth

Location: total_growth.csv

Unlike the previous graphs, these describe top growth in all of France, **minus outliers**. Many of the high growth provinces used to be around 0 revenue, and then were hit with a tax that increased their income significantly. High growth communes tend to be lower income, generally, and so have a greater range of values when showing their income and growth rates. For example, a commune with an income of **€100** that finally gets a normal land tax applied might shoot up to **€100.000**, which is not a lot of money, but still a **100,000%** increase in income.





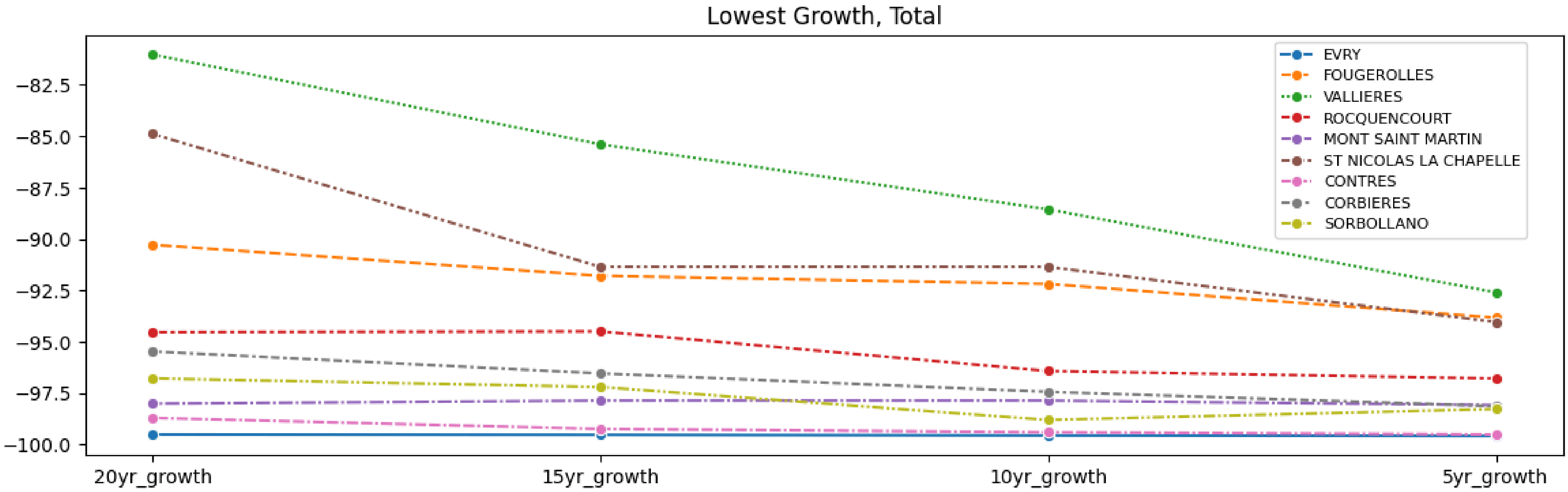
Municipality Rankings - Top / Bottom by Growth

Location: total_growth.csv

Bottom growers change so frequently that it's not worthwhile making a time series graph of every commune, like before. Instead we show the history of lowest growers from the last 5 years, projected out to 20 years in the past.

Importantly, much of the data in this dataset is at **-100%** growth. This is because of missing data, where one year a commune's income dips down to 0. For this graph, we remove such data before visualizing it, although several provinces are still quite low at **-96%** or so. For these cases, it's unclear why the income dropped so much - usually this is the result of a specific tax dropping down to 0, which is more of an issue with local laws than anything else.

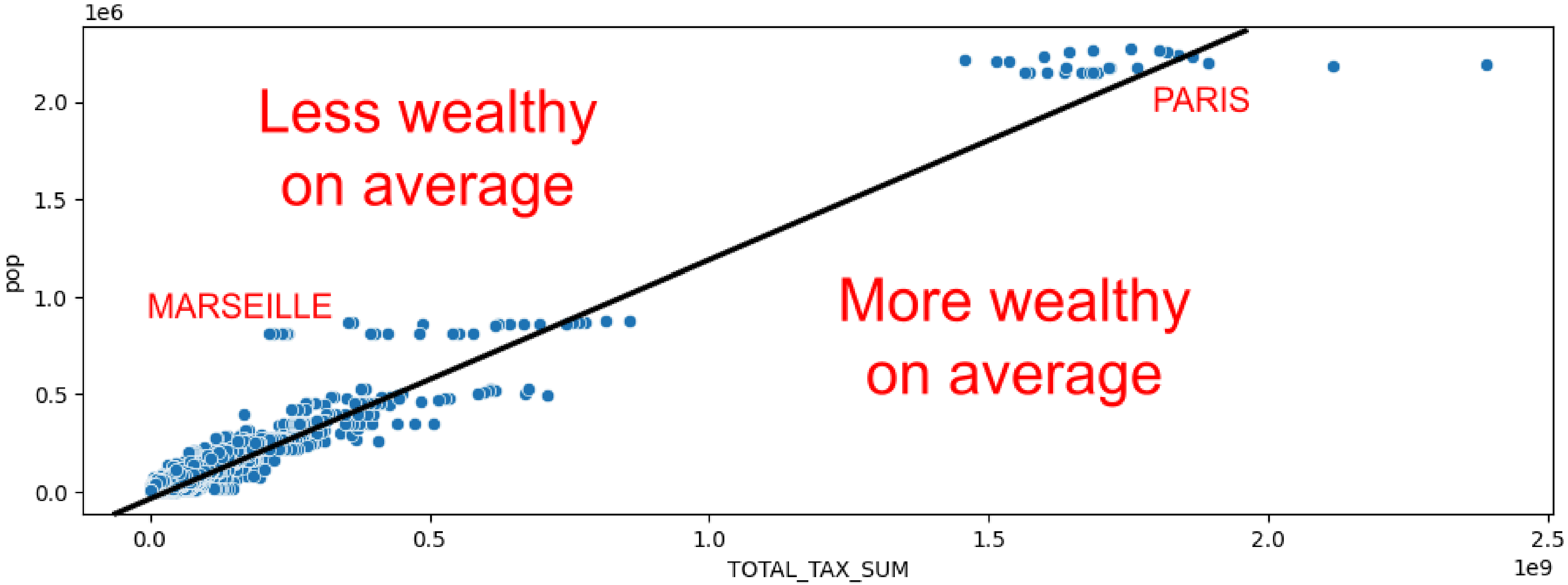
For these straighter lines, this data implies that **FOUGEROLLES**, for example, has seen a consistent 90% decrease in income every single year. It's more accurate to read this data as, **FOUGEROLLES'** income is 90% less than what it was 5 years ago (compared to today), and 90% less than it was 10 years ago (compared to today), etc. So about 5 years ago, its income dropped by 90% (and didnt grow or shrink much before then) and has not recovered.



Correlation Between Population and Revenue

Correlation of **0.97**, meaning that these two variables (population / total revenue) have a very strong relationship. In fact, a **0.97** is about as close as you can get to a perfect relationship with real world data. We explore other variables later to see if we can find a better correlation in our AI models.

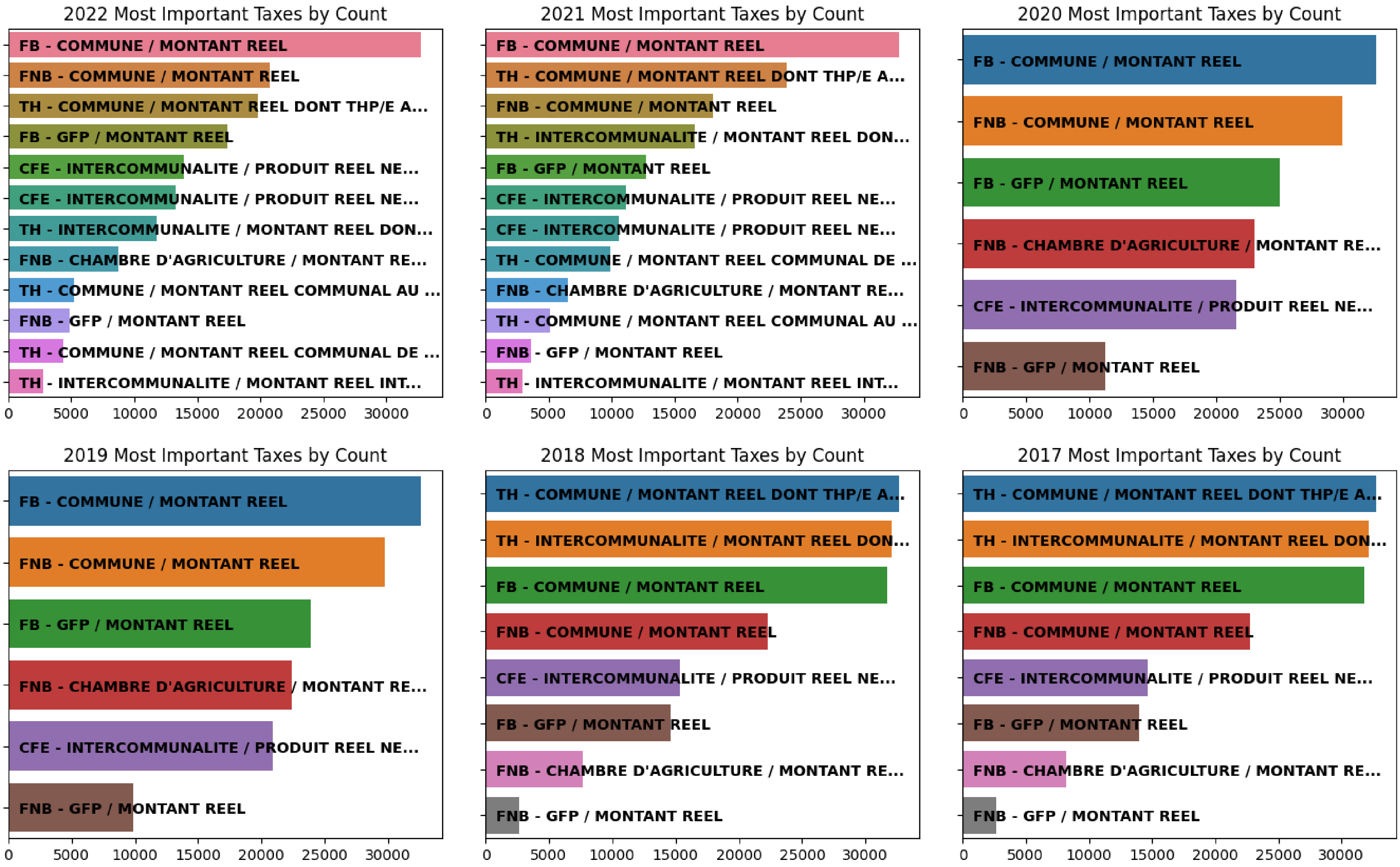
Since we choose not to eliminate outliers in this dataset, you can see several of them below - mostly high population / high income in the top right of the graph. Likely they are communes that contain a skewed amount of extremely wealthy individuals. In this case, the top right outliers are **Paris** and **Ville de Paris**, depending on the time you look at this data.



Significant Taxes

Location: Most_Important_Taxes_By_Count.csv

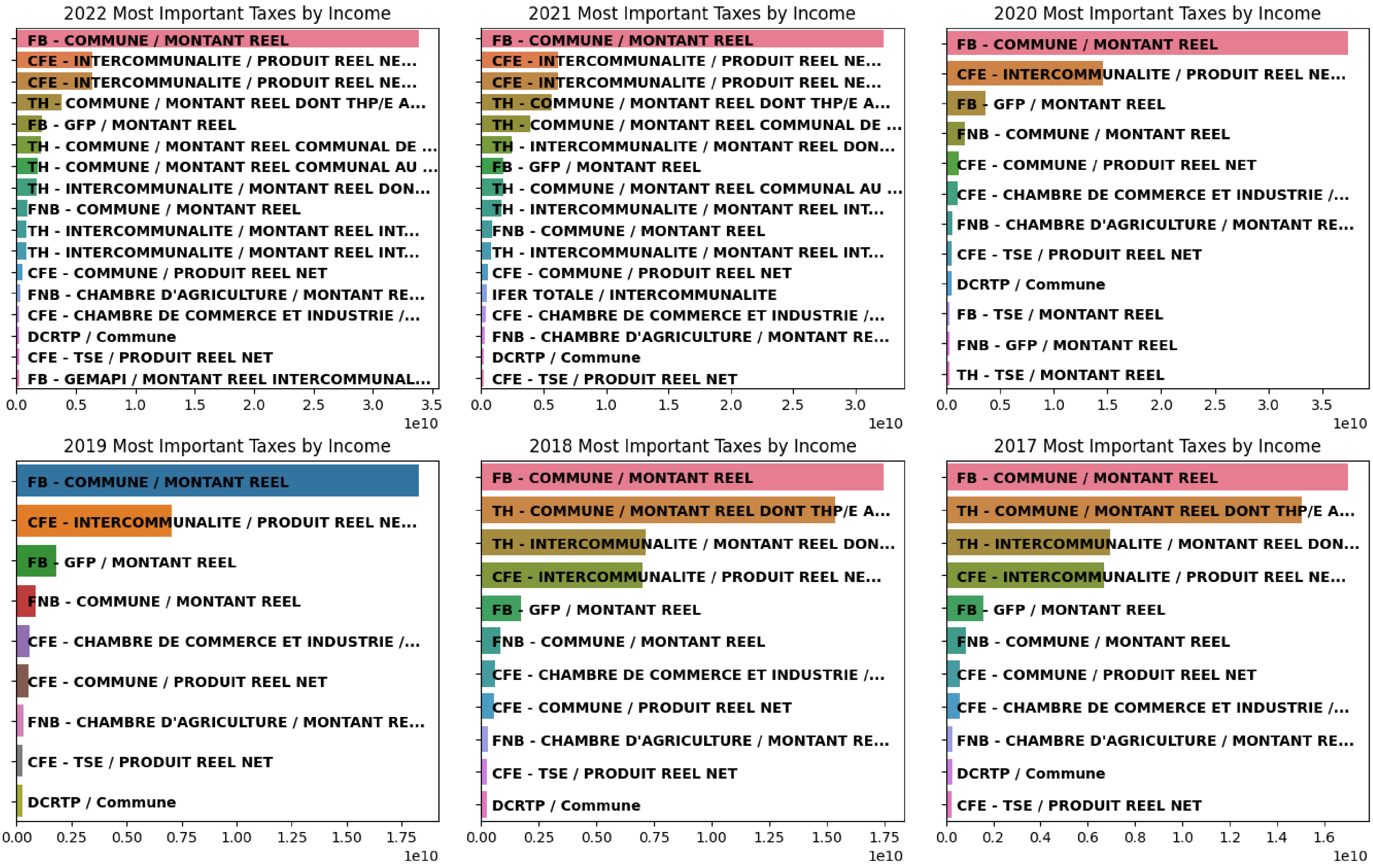
Significant taxes are categorized two ways: **by count**, and **by total income**. *By Count* refers to how often a specific tax appeared in the top 5 taxes of a commune's revenue that year. We can see it changes over time, but **FB** tops from 2019-2022, before that **TH** tops. *By Count* is a good way to determine tax "popularity" instead of looking only at how much a tax may produce.



Significant Taxes

Location: Most_Important_Taxes_By_Total_Income.csv

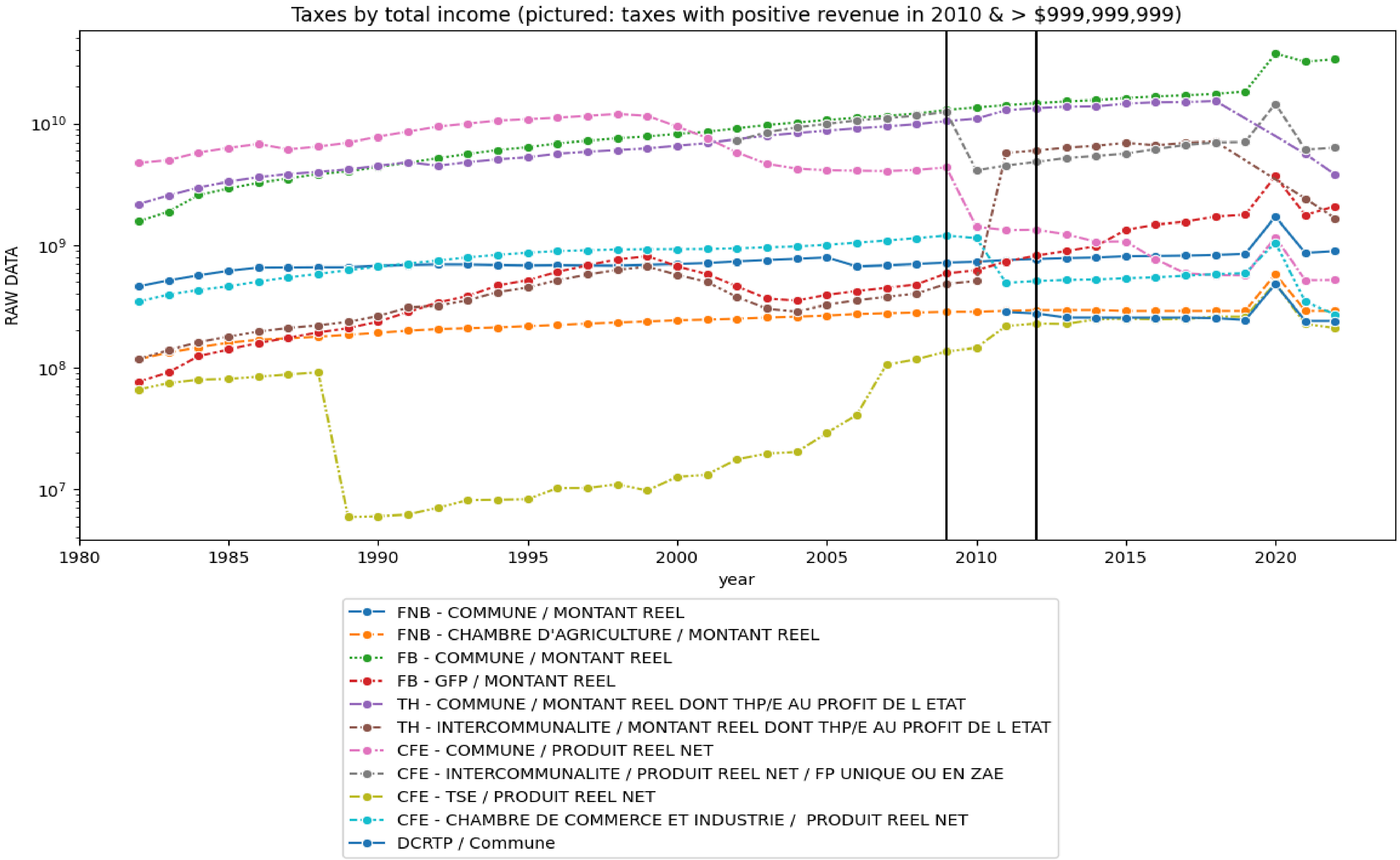
Here we see the same data, but now sorted by total income. It's clear that **FB** surpassed **TH** in total income around 2017 and has continued to outpace other taxes as time moves on, followed by **CFE**.



Significant Taxes

Location: Most_Important_Taxes_By_Total_Income.csv

It is possible to create a time series graph of significant taxes by income, but because there are so many, the graph gets a little messy. Here we see a graph with filtered properties that help us explain the next section, Professional Tax, that includes the top grossing taxes in France. The difference in income sees a slight drop across the board, but is recovered later. We explore this further in the Professional Tax section.

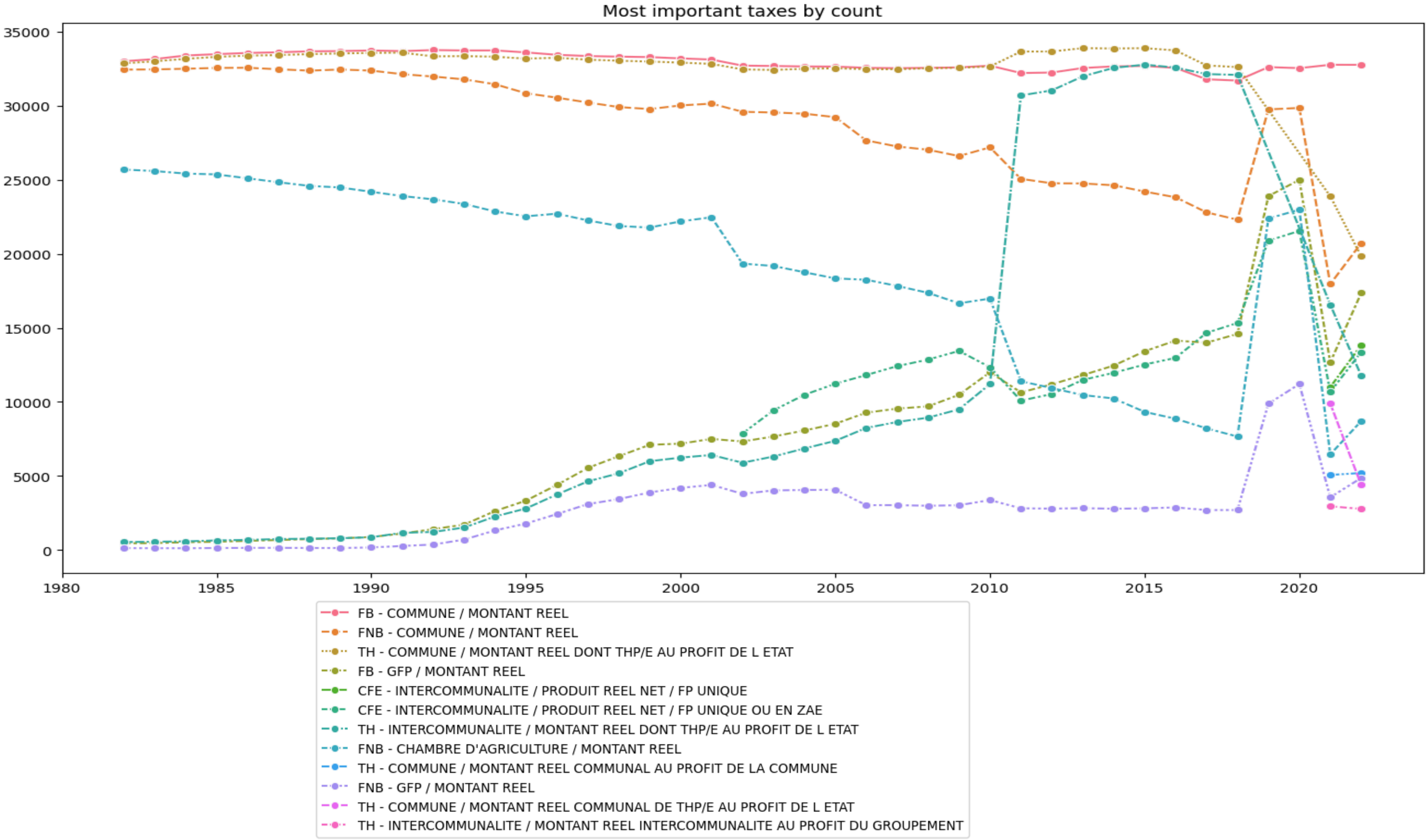


Significant Taxes

Location: Most_Important_Taxes_By_Count.csv

The count graph helps us understand bigger trends a little better too, and evens out the graph a lot. Importantly, it allows us to identify which taxes are more "popular" and then cross reference them with the data above.

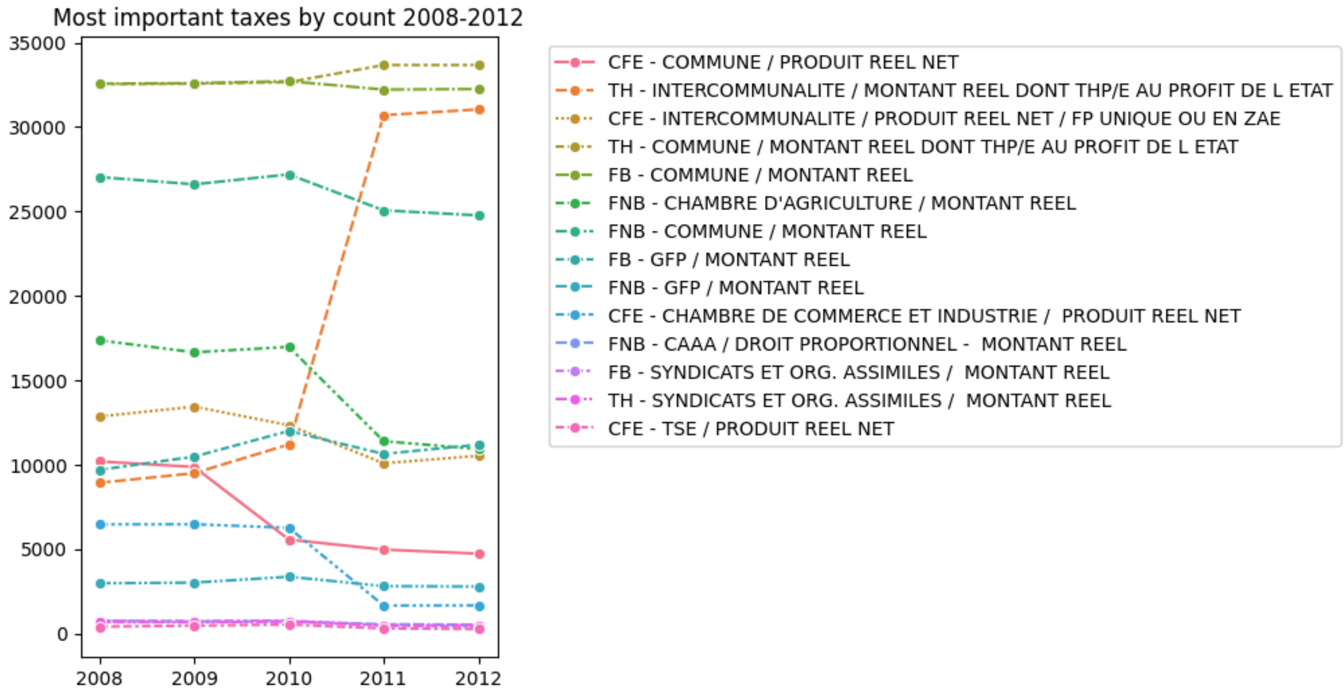
We see a big jump around 2018/2019 and then a big dip in 2021 for many of the taxes here. This jump is not reflected in the total income graph above. This means that during these times, they appeared in the top 5 taxes of a commune significantly more often, suggesting they replaced a different set of taxes in popularity, which according to this graph probably included **TH - INTERCOMMUNALITE [...]**. It's possible the entire TH tax group was affected.

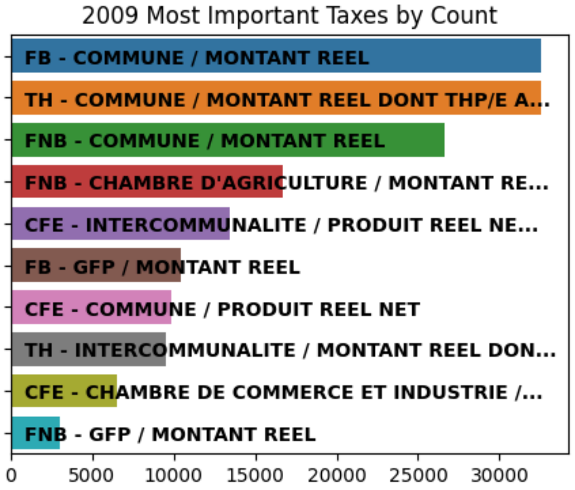


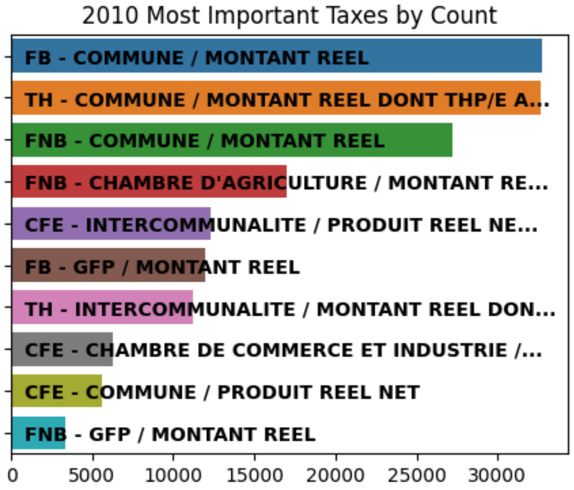
Professional Tax

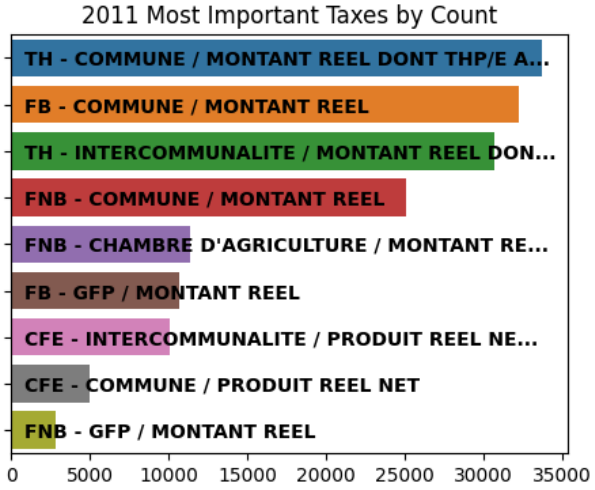
Location: Most_Important_Taxes_By_Count.csv

We fail to identify a single tax that controls the Professional Tax (**TP**), so we turn towards looking at other trends in the data over this time. Notably, we identify the fall of **CFE/FB/FNB** taxes, and the sharp increase in the count of **TH** taxes during the time TP was removed.


Most important taxes by count 2008-2012


2009 Most Important Taxes by Count

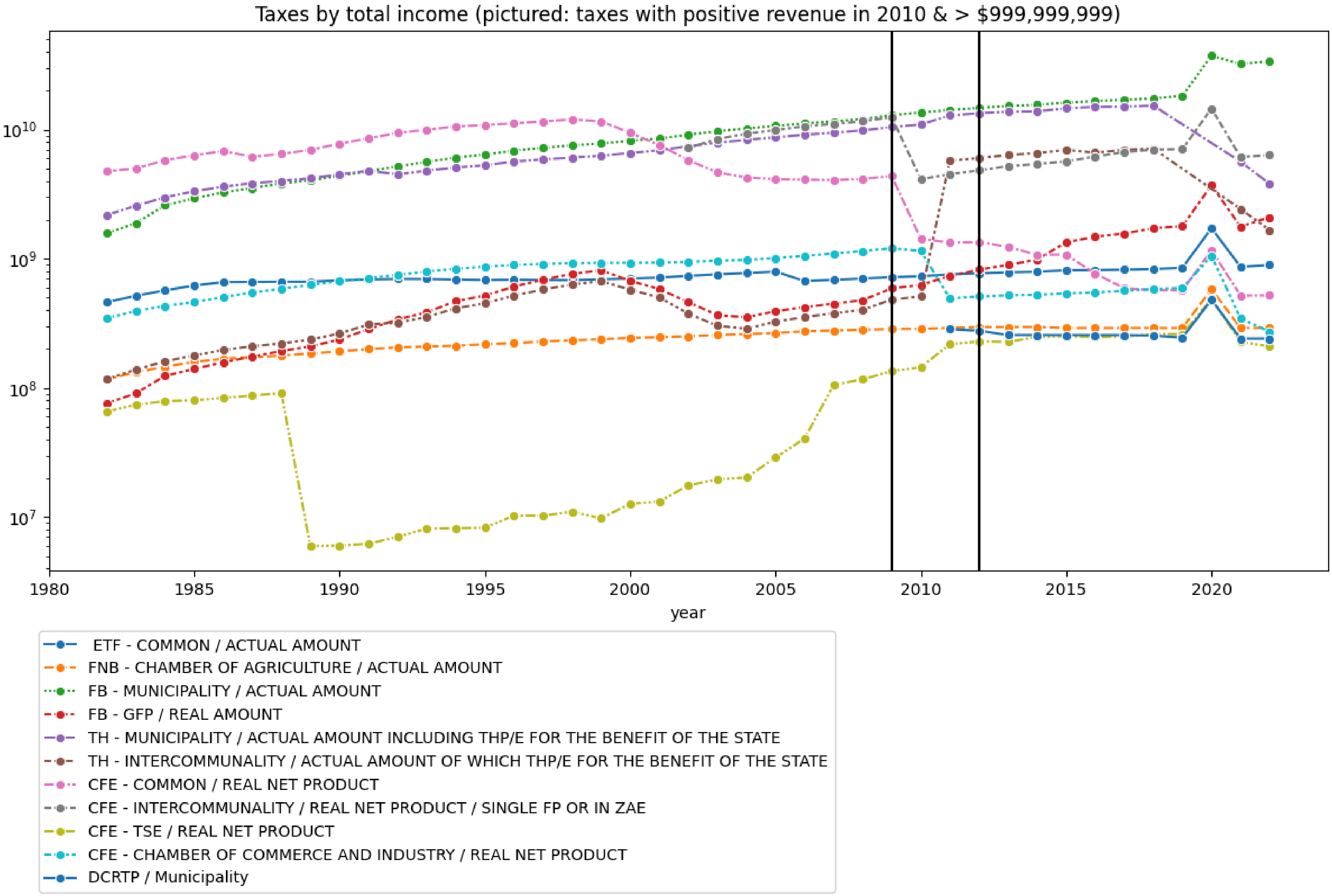

2010 Most Important Taxes by Count


2011 Most Important Taxes by Count

Professional Tax

Location: Most_Important_Taxes_By_Total_Income.csv

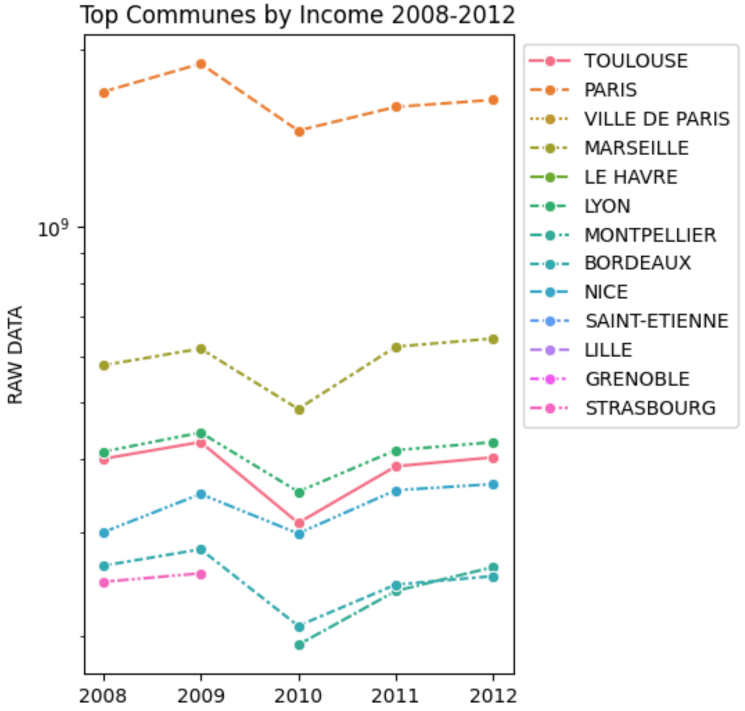
If we look at income instead, the same trend is true. Notably, the decrease in **CFE** / **TH** is larger than the single increase in **TH**.

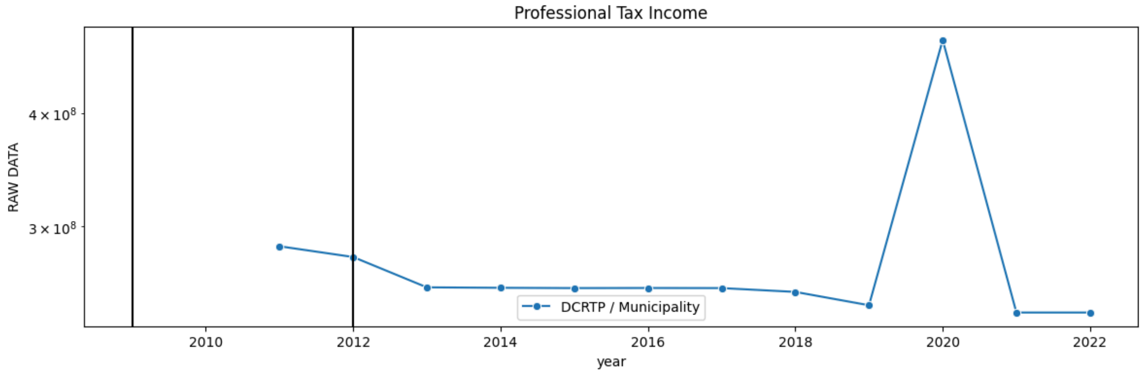


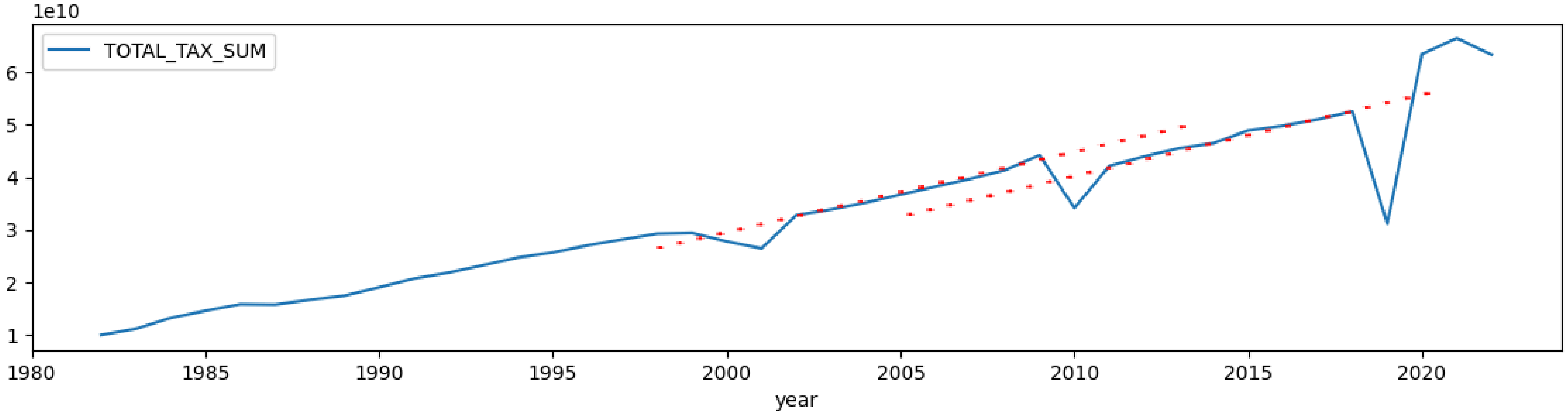
Professional Tax

Location: Most_Important_Taxes_By_Total_Income.csv

Finally, if we look at commune income during this time, we see an obvious dip. Also, the single identified TP tax, **DCRPT**, only appears after 2010, and doesn't make an appreciable difference in income after the fact. We do identify a **global dip in income after 2010, and a normal growth rate following**.







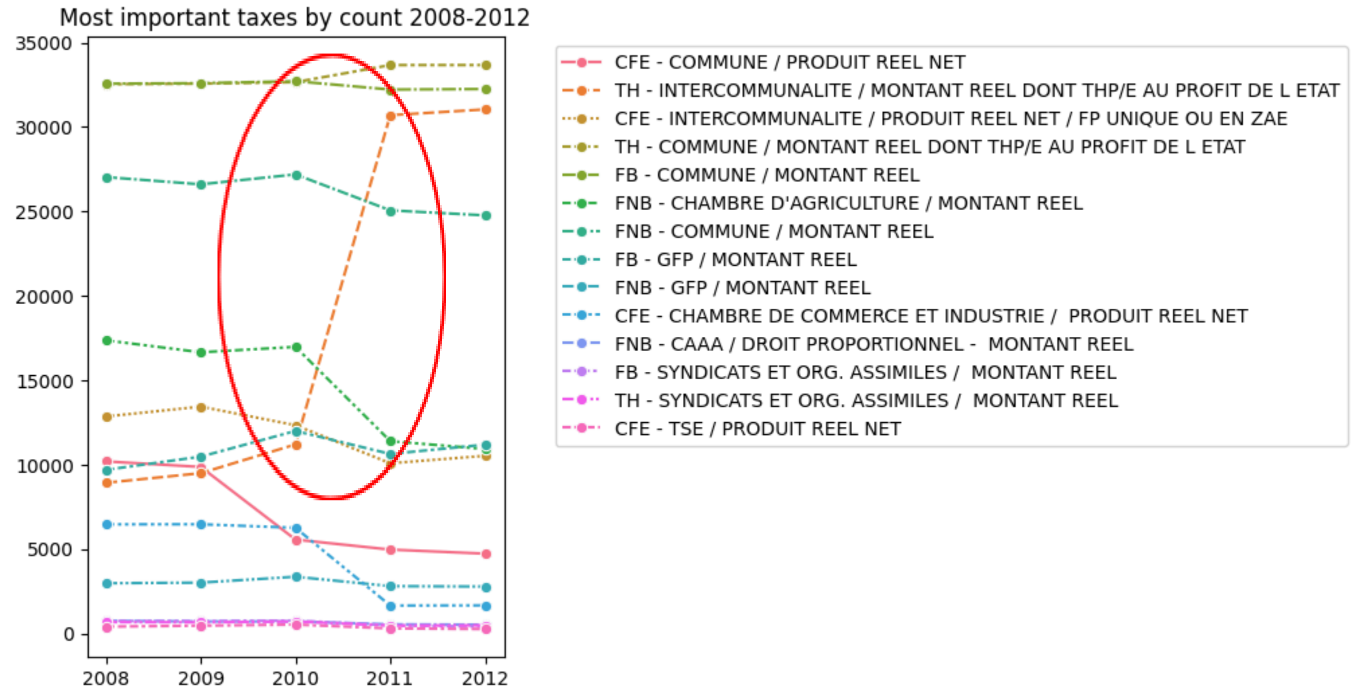
Professional Tax

Key Findings:

- Total income for individual communes in France seems to have decreased slightly after the TP tax change, but recovered after
- Total income for France seems to have decreased slightly because of this tax
    - The rate of change after 2010 did not increase between 2011-2015, **suggesting** that the **removal of this tax did not increase French tax income over time**.
    - If the change was positive, you might expect a drop (as seen) but a faster recovery, and a faster positive growth rate. Instead, **we see a drop, and the growth rate remains unchanged**.
    - **This does not mean such a change was negative -** it's possible that even if income was reduced, GDP increased since theoretically a smaller federal tax income means the population spent less money on taxes, and increased their wealth slightly.
    - It's also likely that such a change may not see a positive return until 10+ years, where we only look at 2-5 years here.
- The TP tax was abolished, but we can see that the DCRTP tax exists after the fact
    - Probably, this tax no longer represents the TP tax as it did before, and has been recategorized into something else
    - We do not have data about the TP tax before 2010 explicitly, only after
- TH and CFE taxes were affected the most between 2010 and 2011
    - **The specific income of some taxes changed** and balanced out, with a **slight decrease** in total income
    - **The abolition of the TP tax was more significant than one tax changing, since several other taxes were significantly affected.**
- Three CFE taxes decreased, one FB tax decreased, and one TH tax increased
- The dip in income in 2010 was caused by the CFE and FB taxes decreasing; whether the TP tax is/was included in CFE before 2011 is unknown.
    - CFE probably contained the TP tax and was then separated into its own tax category after 2010, but that's a question for a tax professional

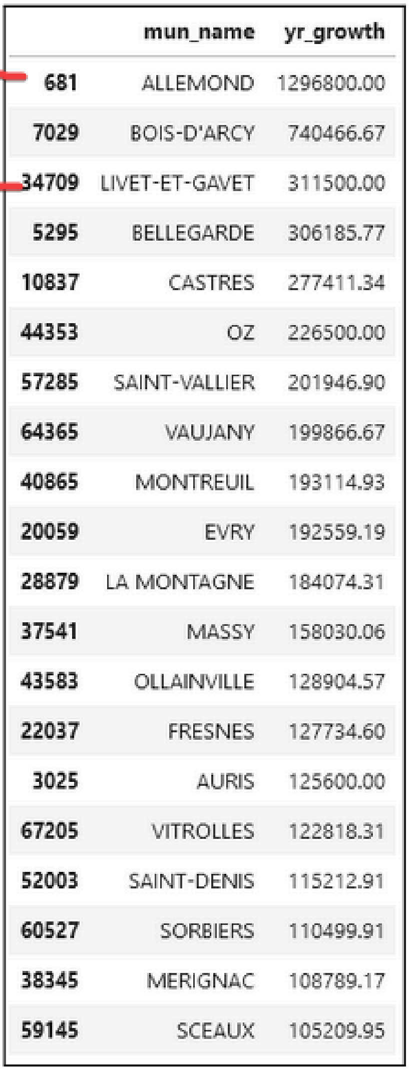
More on the TH Tax

I was curious about the jump located in this graph, so I took a deeper look.

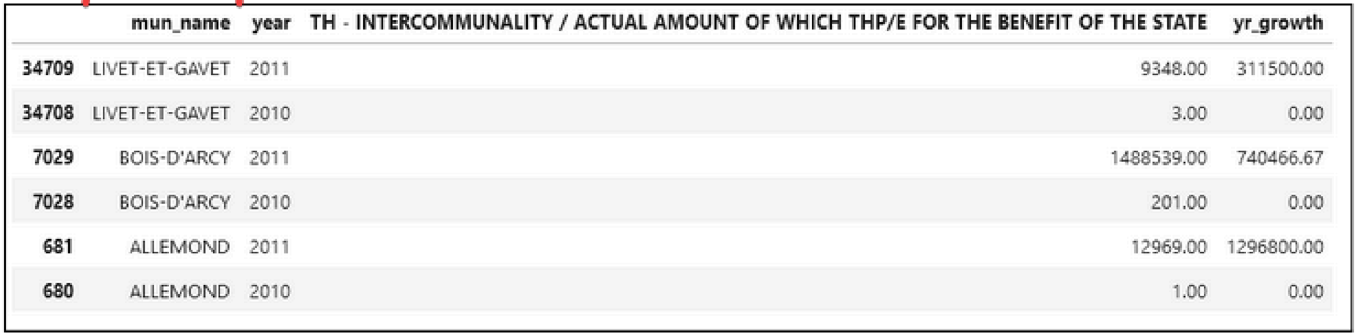


Between 2010 and 2011, which municipalities were affected by the increase of the TH tax the most? On the graph to the right, we can see the answer to that question - most of these top growers, as before, had a very low initial income to start with, which is why their growth percentage is so high.

We can see that below - the tax went from somewhere around single digits to over 5 - hence the extreme increase.



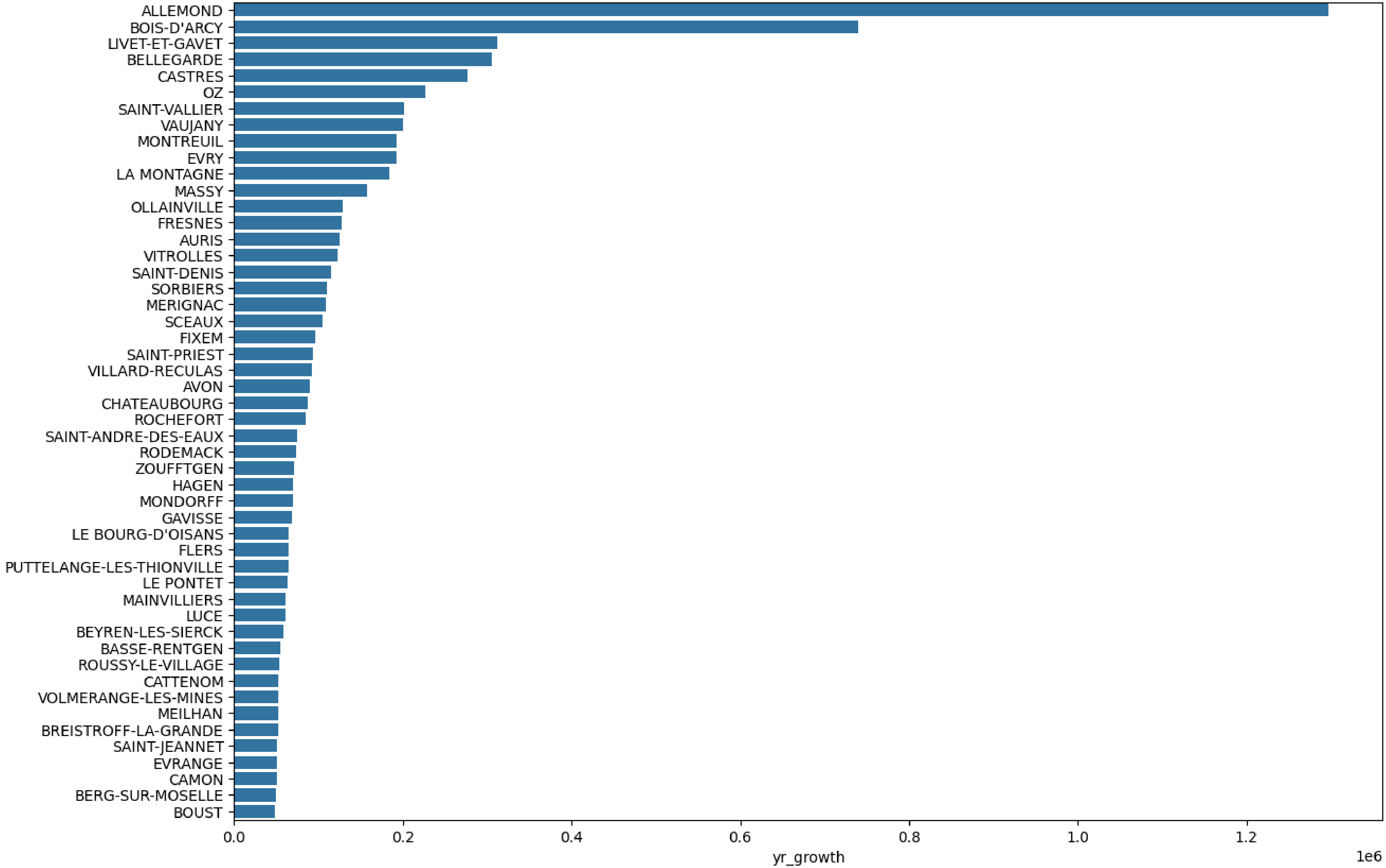
	mun_name	yr_growth
681	ALLEMOND	1296800.00
7029	BOIS-D'ARCY	740466.67
34709	LIVET-ET-GAVET	311500.00
5295	BELLEGARDE	306185.77
10837	CASTRES	277411.34
44353	OZ	226500.00
57285	SAINT-VALLIER	201946.90
64365	VAUJANY	199866.67
40865	MONTREUIL	193114.93
20059	EVRY	192559.19
28879	LA MONTAGNE	184074.31
37541	MASSY	158030.06
43583	OLLAINVILLE	128904.57
22037	FRESNES	127734.60
3025	AURIS	125600.00
67205	VITROLLES	122818.31
52003	SAINT-DENIS	115212.91
60527	SORBIERS	110499.91
38345	MERIGNAC	108789.17
59145	SCEAUX	105209.95



	mun_name	year	TH - INTERCOMMUNALITY / ACTUAL AMOUNT OF WHICH THP/E FOR THE BENEFIT OF THE STATE	yr_growth
34709	LIVET-ET-GAVET	2011	9348.00	311500.00
34708	LIVET-ET-GAVET	2010	3.00	0.00
7029	BOIS-D'ARCY	2011	1488539.00	740466.67
7028	BOIS-D'ARCY	2010	201.00	0.00
681	ALLEMOND	2011	12969.00	1296800.00
680	ALLEMOND	2010	1.00	0.00

More on the TH Tax

More context - which regions were affected the most by this tax jump?



AI Models

We can create a model to predict the following:

Total Income Regression
We identified earlier that there is a very clear relationship between population and total income. We also assume that a clear linear relationship exists with every tax and a total of those taxes, since it's just addition. If we did this, we could use either just population, or population and all of the applicable taxes that predict total income for a commune.

Specific Tax Regression
Like total income, we could also do it with specific taxes and see how they may change over time.

Classification
There are a good many variables in this dataset that fall into a classic classification problem. For example Variable 1091 is a yes/no variable that describes if *"The municipality is partially or totally within a regional aid area"*, or Variable 1107: *"The municipality has instituted an increase in council tax on second homes"*. Ultimately these could be used to answer a question like, "**Does X commune fall into the high tax bracket?**" or "**Is X commune part of Y region?**". The problem with both of these, and any derivatives, are that we already have this information and need not predict such features; "tax brackets" could be answered with a simpler algorithm (or a calculator) and we already know what communes are in each region.

To keep things simple, we'll go with a **Linear Regression,** since our data is pretty linear anyway and has a high correlation, and keep it to predicting the total tax sum for regions. We might consider using a different type of regression model if our data had a lower correlation (say 0.6 or lower), but a **0.97 correlation suggests the data is very linearly related.**

We identify two top models with a **95.16%** accuracy, and a **94.62%** accuracy. We decide to use linear models because our data is very linear (as you tend to find with addition - there is a perfectly linear relationship between components of a sum, and their sum), however we identify that population ends up being a better predictor of province income over time, or at least an approximately equal predictor when compared to using all of the individual taxes of a region (and not their populations).

In practice, the accuracy of these models is less than 1% difference, and the difference can be attributed to random chance. With different parameters, one may be better than another, so it's hard to say which one is really better. Because they are so close, we will conclude that **they are similar enough to both be accurate predictors**.

Interestingly, the Linear model with all taxes (and not population) had an **82.36%** accuracy, which we may not expect because it's just addition.

Thanks for reading!

If you have any questions, please reach out to https://github.com/MyNameIsCalvinDavis/

Appendix

Find specific file downloads here: https://github.com/MyNameIsCalvinDavis/Data-Science/tree/main/French%20Tax%20Data

Does not include the parquet / dataset file, as it is too big. Reach out to the **Joint Research Centre (JRC) / Desights.ai** for access to this file, or Christian Casazza (https://github.com/ChristianCasazza/datasharing) who provided the API to access this data, and is available on github. The bottom portion of the python notebook provides instruction for how to download the dataset.

Running the attached python notebook file (data.ipynb) will automatically produce all of these files, but I provide them to download separately. The only file not included is *RAW_DATASET_name_totalsum*, which is too big to upload, but also generated upon execution.

Most_Important_Taxes_By_Count.csv

- *Provides information about specific taxes and their frequency in the top 5 taxes of a commune. Missing values mean that tax did not show up in the top 5 for any commune that year.*

Most_Important_Taxes_By_Total_Income.csv

- *Provides information about specific taxes and their total income over time. Missing values mean that tax had no information for that year.*

summary_summary.csv

- *Provides summary statistics of the dataset, including all taxes and total income. Does not include data about specific communes.*

Total_Growth_of_Communes_TH_INTERCOMMUNALITY_Tax.cs

- *Provides income growth information for communes regarding this specific tax, since it jumped up the most after 2010*

total_growth.csv

- *Provides income growth information for all communes for 5 / 10 / 15 / 20 years. TOTAL_TAX_SUM in this spreadsheet is for 2022.*

translated_cols.csv

- *Originally used to help me understand column names, since I don't speak french. I threw all the columns into google translate and this is the result.*

fr_cols.txt

- *The column names taken directly from the database file*

data_descriptions.xlsx

- *Information about the dataset provided by the original contest owner*