Assignment 1: Data and Descriptive

Anine Therese Karlsen & Mona Lisa Jones

This assignment aims to acquire, process, and analyze sub-national GDP and population data for a subset of European countries. Calculate GDP per capita and explore regional inequity using various descriptive statistics and visualizations.

1 Introduction

While national GDP and GDP per capita are vital indicators of a country's aggregate economic health, they do not shed light on how wealth or income is distributed among its residents. A high national GDP can, paradoxically, coexist with pockets of regional deprivation (**lessmann2017?**).

The truth of this statement becomes more evident when taking a closer look in to into sub-national data. Regional wealth disparities are of prime concern, especially when crafting policies for equitable growth (**lessmann2017?**). A country's macro-level prosperity does not automatically guarantee that all its regions partake equally in this wealth. By studying smaller regions within a country, it is possible to get a more nuanced narrative about the state of regional economic disparities (**lessmann2017?**).

This assignment is anchored in this very premise. It seeks to acquire, process, and analyse sub-national GDP and population data for a selected subset of European countries, namely France, Denmark, Hungary, Portugal, and Slovakia, spanning the period from 2000 to 2020. The overarching aim is to calculate GDP per capita for these regions and delve into the intricacies of regional inequity. This exploration will be underpinned by various descriptive statistics, visualizations, and analytical techniques to paint a comprehensive picture of regional economic landscapes.

Using time series data observing the GDP per capita and GINI trends across regions and over time, we can discern whether certain areas are ahead or behind in economic performance.

2 Literature review

"In recent decades, the regional distribution of incomes within countries has attracted considerable interest among academics and policy makers." (**lessmann2017?**)

Slovakia, experienced significant economic growth. Their GDP per capita grew faster than the older EU member states, albeit from a lower starting point, leading to a convergence (**fullart?**).

The global financial crisis and subsequent European debt crisis had significant impacts on European economies. GDP per capita declined in several countries, particularly those hit hardest by the crisis, such as Greece, Spain, and Portugal. Some regions took many years to recover to pre-crisis levels, while others, especially in Northern Europe, recovered more quickly (**lewis?**).

Major metropolitan areas, like Paris in France, London in the UK, and Berlin in Germany, often grew faster than other regions in their respective countries, leading to increasing spatial disparities (**kiuru2019?**).

Denmark maintained relatively high levels of GDP per capita throughout this period, benefiting from a diversified economy, strong institutions, and a high degree of economic openness. It did face challenges during the 2008 economic crisis but recovered relatively quickly (**dynamics?**).

Slovenia, which joined the EU in 2004, showed growth in GDP per capita, though it too was affected by the global economic downturn in 2008. However, it managed to recover in the subsequent years (**fullart?**).

Regional disparities persisted, with the Île-de-France region (which includes Paris) significantly outperforming other regions in terms of GDP per capita. The gap between urban and rural areas also remained a topic of discussion (**regional?**).

Towards the end of this period, in early 2020, the COVID-19 pandemic posed a new set of challenges for European economies, causing significant contractions in GDP per capita across many regions (**regional?**).

3 Part A: Sub-national GDP and GDP per Capita

3.1 Data Acquisition and datasets

The first thing we did in this assignment was to acquire our data. Trough Eurostat, we could download the datasets nama_10r_3gdp and demo_r_pjanggr3 as csv files, as well as filter the data by our preferences before downloading it. As per the assignment, we filtered the dataset by choosing the years 2000 to 2020. Furthermore, we selected the NUTS 3 region for the nations we were given, which were Portugal, France, Hungary, Slovakia and Denmark. Finally, we specified the data to be in million Euro.

3.1.1 GDP (nama_10r_3gdp)

The nama_10r_3gdp dataset from Eurostat provides insights into GDP at regional level using the NUTS classification system. It furnishes GDP values in both current prices and adjusted for inflation, with some figures given in purchasing power standards (PPS) to account for price level differences between countries. The data is often structured by year and region Eurostat (2023a).

The GDP at market prices represents the final result of production activities of resident producer units within a region or nation. It is calculated as the sum of the gross value added across various institutional sectors or industries, augmented by taxes and reduced by subsidies on products (which are not allocated to specific sectors or industries). This also balances out in the total economy production account. In terms of methodology, while national accounts compile GDP from the expenditure side, regional accounts don't adopt this approach due to the complexities of accurately mapping inter-regional flows of goods and services.

The different measures for the regional GDP are absolute figures in € and Purchasing Power Standards (PPS), figures per inhabitant and relative data compared to the EU Member States average Eurostat (2023c).

3.1.2 Population (demo_r_pjanggr3)

Using the NUTS categorization once more, Eurostat's demo_r_pjanggr3 records annual population changes at the regional level. This dataset includes information on births, deaths, net migration, and may also include demographic information on age and gender. It's also often displayed in a year-by-region format, and the data usually spans in yearly intervals Eurostat (2023b).

Eurostat's primary source for yearly demographic data at the regional level stems from the Unified Demography (Unidemo) project. The project covers 37 countries and is the central repository for demographic and migration-related data. Specific metrics gathered under UNIDEMO encompass population counts at the close of the calendar year and events such as births and deaths occurring within that year. Additionally, data on marriages, divorces, and migration flows are recorded.

For the purpose of this research, the demographic data references the NUTS 2016 classification, which provides a detailed breakdown of the European Union's territory Eurostat (2021).

3.1.3 NUTS classification

The Nomenclature of Territorial Units for Statistics (NUTS) offers a stratified system to segment the economic territory of the EU (including the UK) to facilitate the consistent collection and harmonization of regional statistics across Europe. The NUTS regions range from NUTS 0 Country level to NUTS 3 small units such as municipalities level.

3.2 GDP per Capita Calculation

The formula for calculating GDP per Capita is as follows:

$y_i = GDP_i/population_i$

After calculating the GDP per capita for all NUTS 3 regions in our assigned countries, we can see that there is a large spread between the figures for the various regions. In this assignement we want to look at regional inequity; in order to do this in a valuable way we have to divide between the different countries. By doing this, we can gain important insights on regional differences that we can utilize, for instance, to discuss national policy on equity and sustainable economic development in regions.

3.2.1 Descriptive statistics

In this part we will report and interpret different types of essential descriptive statistics. Measuring regional income inequality is challenging due to heterogeneity of regions. The number of regions in our data set varies largely in size and population. Since the focus of this paper is purely growth and changes in inequities over time, the variations of size and population density becomes a minor issue because the country-level territorial heterogeneity is fixed.

"Interest in income inequality has led to the development of several ways of measuring it. Two types of measures are of interest in this paper—-static and dynamic. *Static measures* provide a snapshot of these inequalities at a point of time whereas the dynamic *measures capture historical trends.*" (læreboken)

In this part, we'll look at GDP per capita for our assigned countries on a NUTS 3 level. In addition, we'll use different kinds of descriptive statistics in order to further 0,37, is analyse this data. In this analysis, we'll use Wooldridge (2020) for help.

By using figures, we can visualize the GDP per capita, and look at how it varies among the different regions. In these figures, a line represent one NUTS 3 region.

Mean

Calculate the mean to provide a representative value for a dataset, facilitating understanding of its central tendency and serving as a benchmark against which deviations and anomalies can be assessed, in later steps when building and interpreting regression models.

MMR

A comparison of the GRDP (gross regional domestic product) per capita of the region with the highest income to the region with the lowest income (minimum per capita GRDP) provides a measure of the range of these disparities. If this measure is small (close to 1), then it would mean that the different regions have relatively equal incomes. If this measure is large, then the interpretation is more problematic, as it does not tell us if the high ratio is due to substantial variation in the distribution of per capita GDRPs or the presence

of outliers. Nevertheless, maximum to minimum ratio (MMR) provides a quick, easy to comprehend, and politically powerful measure of regional income inequality.

Standard diviation (SD)

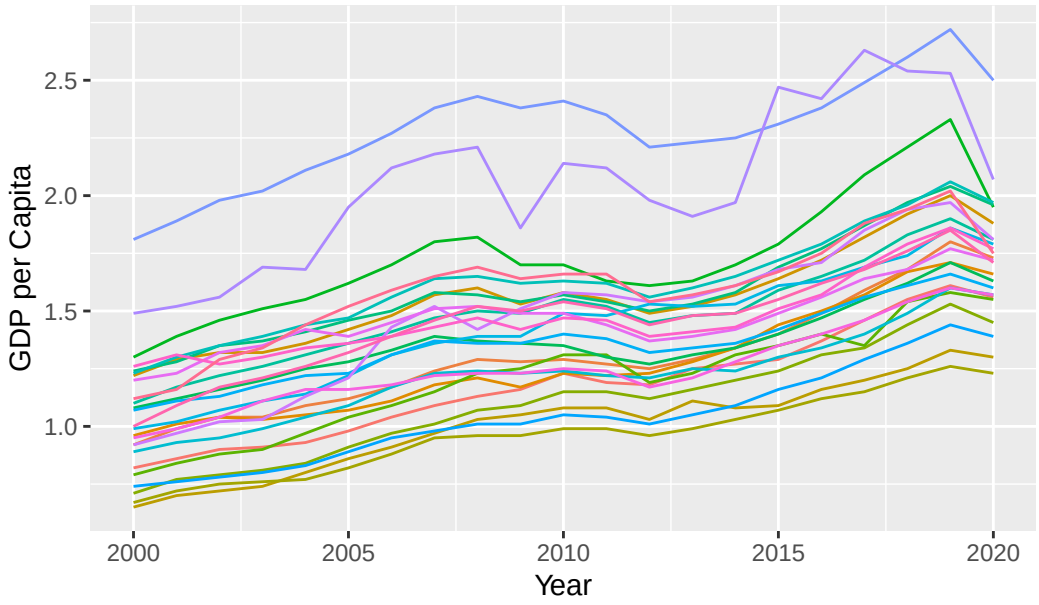
Calculating SD to quantify the dispersion or variability of a data set around its mean. Helping us assess the degree of uncertainty, variability, or risk associated with an economic variable or parameter, which is crucial for understanding the reliability of estimations and predictions (Wooldridge, 2020).

Median

The median serves as a robust measure of central tendency, especially when a dataset may have outliers or is skewed. Unlike the mean, the median is not influenced by extreme values and, thus, can provide a clearer picture of the "typical" value in situations where the data distribution is not symmetrical (Wooldridge, 2020).

3.2.2 Portugal

Figure 1: GDP per Capita for Portugal

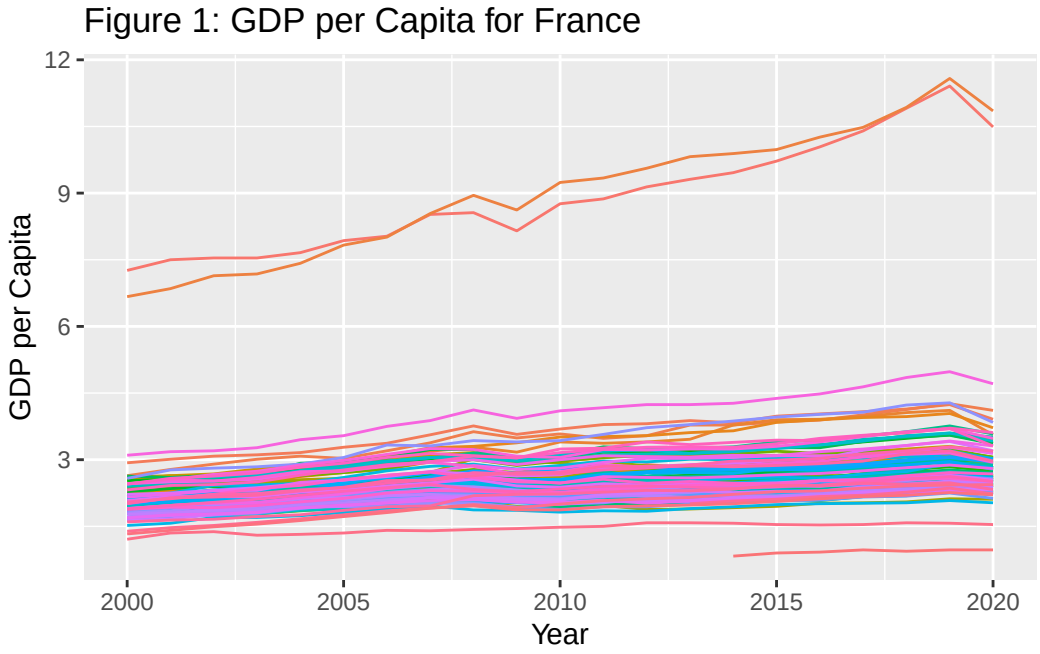


	GDP_per_capita
mean	1.4185524
median	1.3900000
std_dev	0.3702905
minimum	0.6500000
maximum	2.7200000

By looking at figure for Portugal, we can see that the GDP per capita in Portugal's regions appears to be fairly consistent. There is however some regional variability. We can see that the regions around the big cities like Lisbon have a higher GDP per capita compared to some more rural areas. Since Lisbon is the capital of Portugal, there is probably a higher concentration of industries, making it a economic center (which again makes the GDP per capita higher).

To continue, we can see that the mean is a little higher that the median, something that might indicate that regions like Lisbon are pulling up the average. If we compare the standard derivation for Portugal with the other countries, we'll see that is fairly low in comparison. This might mean that there is not a lot of variability between the GDP per capita across different regions in Portugal. The gap between minimum and maximum is also low compared to other countries, something that'll also show us that the economic disparity in Portugal might not be as high as it is in other countries.

3.2.3 France



	GDP_per_capita
mean	2.630444
median	2.440000
std_dev	1.053767
minimum	0.830000
maximum	11.580000

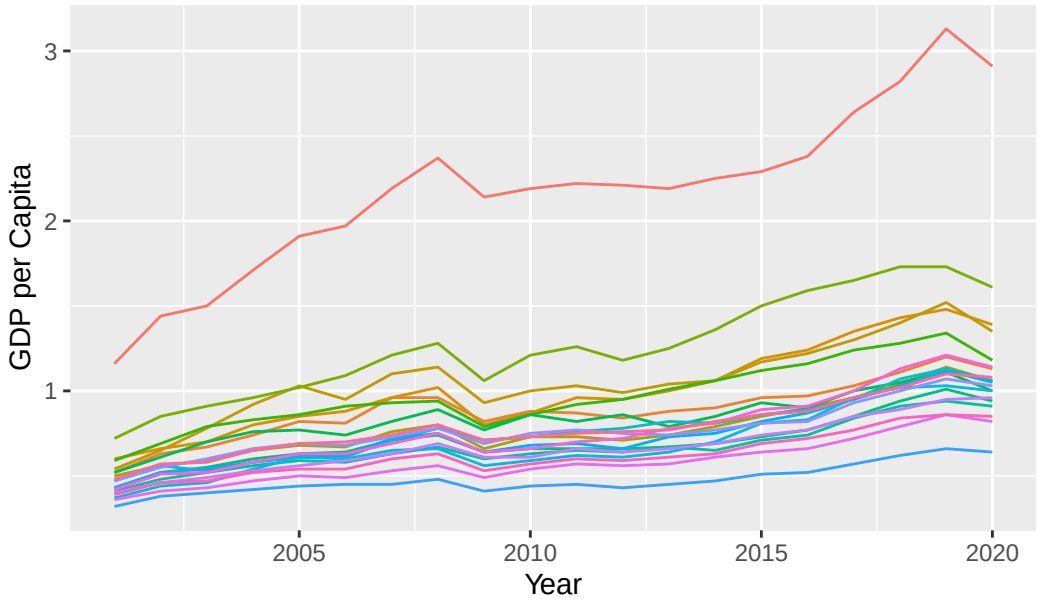
When looking at the figure for France, we can right away see that there are some regions that have a much higher GDP per capita compared to the other regions. The region that has

the highest GDP per capita for all years is the île-de-France region, one that also includes Paris. This significant difference between the regions with the highest GDP per capita and the lowest, shows us that there is a high concentration of economic activity and wealth in a few urban regions. Similar to Portugal, we can also again see that there is a difference between urban and rural regions.

Just as in Portugal, there is also a higher mean in France as well. Something that is different from the data in France compared to Portugal, is that the standard derivation is higher, and the difference between minimum and maximum is large. This strengthens what we have look at earlier in the figure, with some regions having a high concentration of wealth.

3.2.4 Hungary

Figure 1: GDP per Capita for Hungary

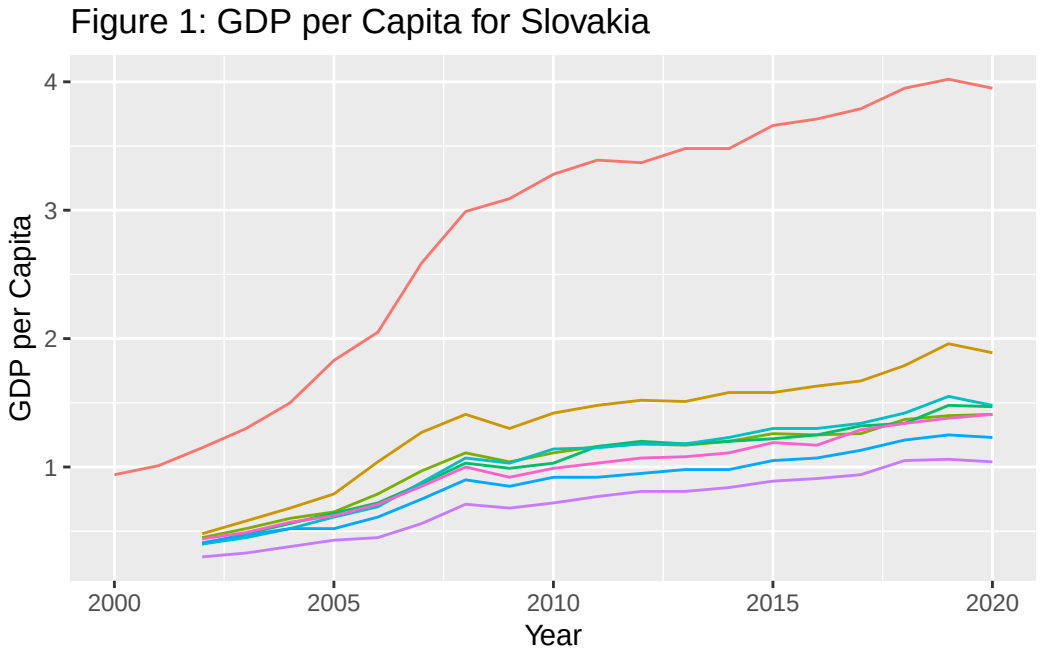


	GDP_per_capita
mean	0.8598000
median	0.7650000
std_dev	0.4059723
minimum	0.3200000
maximum	3.1300000

In Hungary, most of the regions have similar GDP per Capita. One region that sticks out by having a higher value, is the region of Budapest, Hungary's biggest city.

We have here as well an mean that is larger than the median, high standard derivation, and a large gap between minimum and maximum.

3.2.5 Slovakia



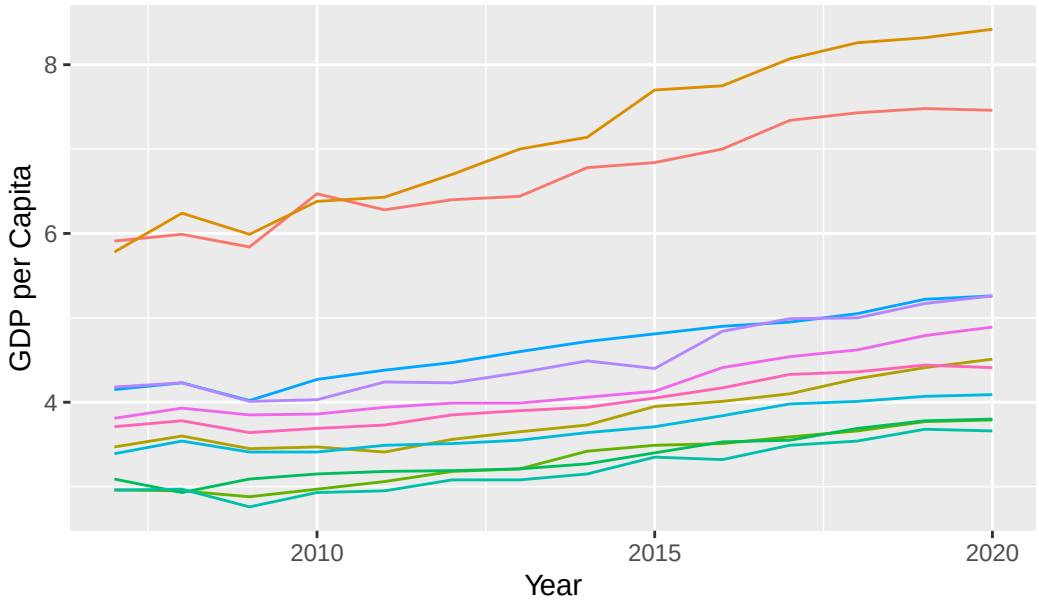
	GDP_per_capita
mean	1.2501948
median	1.0950000
std_dev	0.8018259
minimum	0.3000000
maximum	4.0200000

In Slovakia as well, we have one region that has a much higher GDP per capita than the rest of the regions. This region is Bratislava, which is the biggest city and capital, something that might point to this city being the economic capital of Slovakia as well.

Slovakia has also a mean higher that the median, and a large gap between minimum and maximum. In addition, the standard derivation is pretty high, meaning that there is some regions (or one region in this case) that is far away from the rest of the regions when it comes to economic development.

3.2.6 Denmark

Figure 1: GDP per Capita for Denmark



	GDP_per_capita
mean	4.419221
median	4.000000
std_dev	1.343933
minimum	2.760000
maximum	8.420000

Lastly, we have Denmark. We can see similar pattern here as well, with the capital Copenhagen being one of the regions with the highest GDP per capita.

We can also see the same as the previous countries, with the mean being higher than the median, which shows us that regions like Copenhagen might drag the mean up by being much larger than the rest of the regions.

4 Part B: Regional Inequity

4.1 Gini Coefficient Calculation

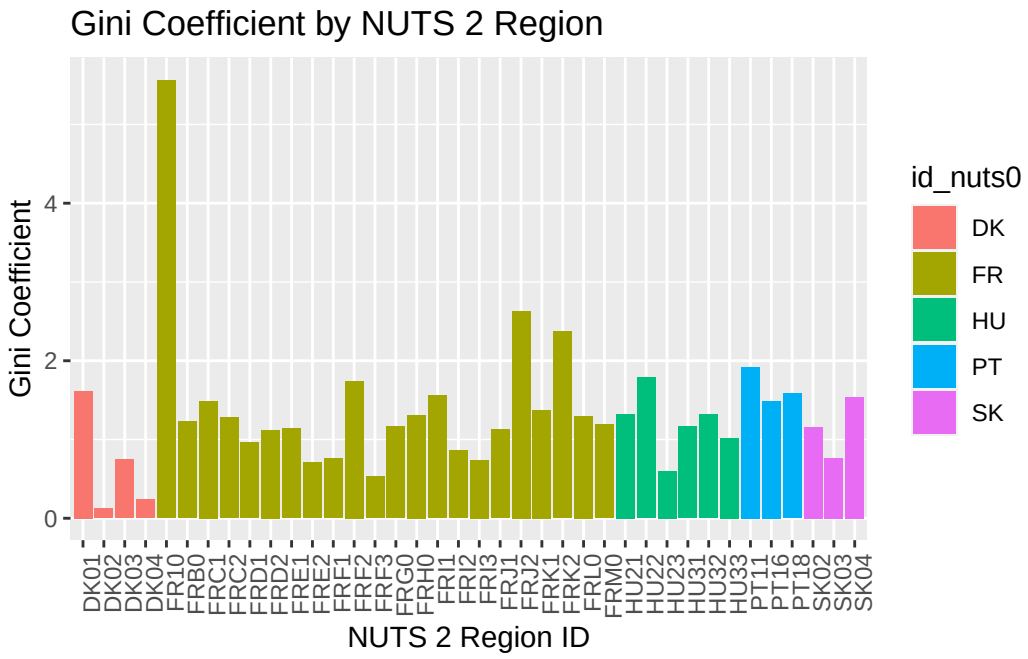
With the use of the NUTS3 GDP per capita data and this formula:

$GINW_j = \frac{1}{2\bar{y_j}} \sum_i^{n_j} \sum_l^{n_j} \frac{p_i}{P_j} \frac{p_l}{P_j} |y_i - y_l|$

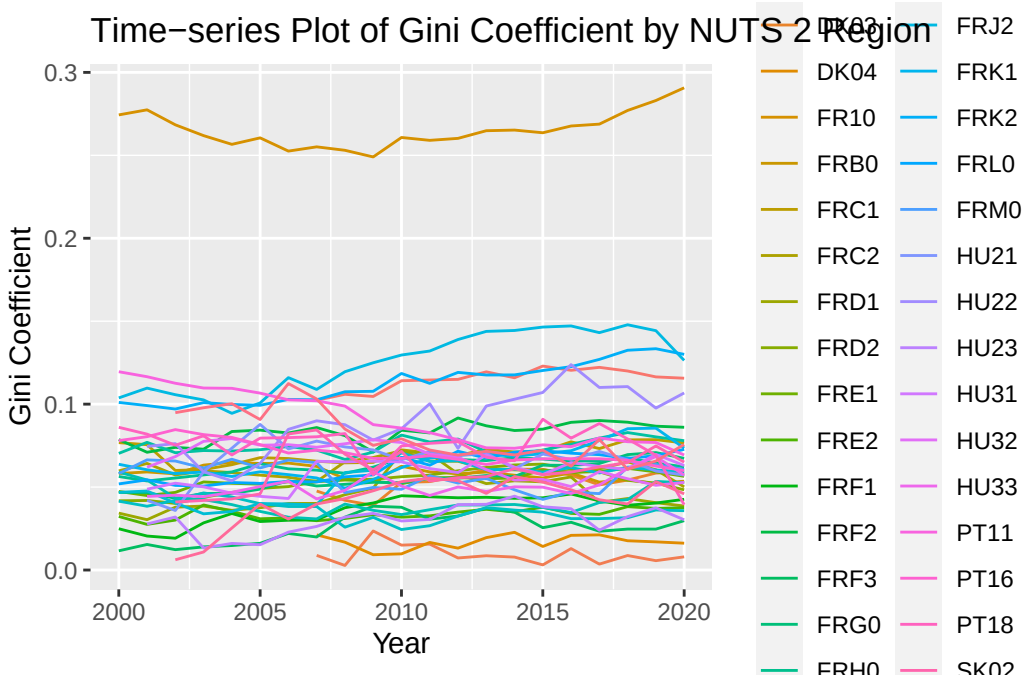
we will compute the population-weighted GDP Gini coefficient for each European NUTS2 region in our assigned countries.

The gini coefficient can help us measure inequality in a distribution, as is therefore a useful tool for us to use when we look at regional inequity. The closer the gini coefficient is to 1, the bigger the inequality is; a number closer to 0 equals equality. When looking at the gini coefficient for NUTS 2 regions, we also get a better overview over differences in income between different regions, and it also makes it easier to find the reasons as to why there is a difference between the regions Hasell & Roser (2023).

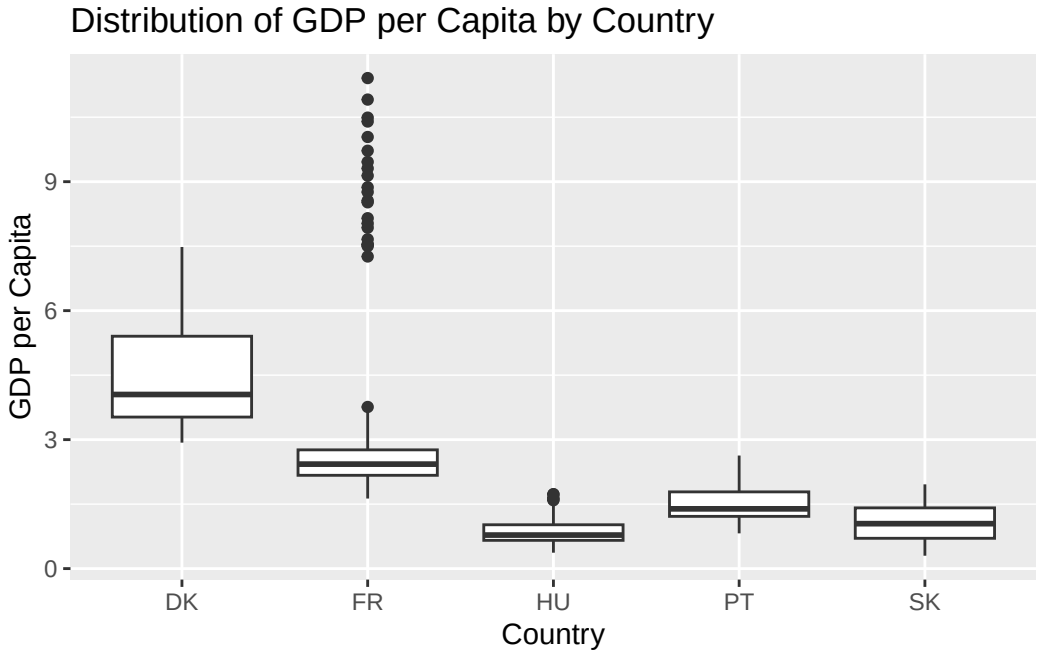
After calculating the gini coefficients, we can see that there are some similarities to the data we got from GDP per capita for NUTS 3 regions. In order to see these similarities better, as well as look for other important aspects that can be provided trough the calculations, we will visualize the data in three different ways.



...


Time−series Plot of Gini Coefficient by NUTS 2 Region
Gini Coefficient
0.3
0.2
0.1
0.0
2000
2005
2010
2015
2020
Year
DK03
DK04
FR10
FRB0
FRC1
FRC2
FRD1
FRD2
FRE1
FRE2
FRF1
FRF2
FRF3
FRG0
FRH0
FRJ2
FRK1
FRK2
FRL0
FRM0
HU21
HU22
HU23
HU31
HU32
HU33
PT11
PT16
PT18
SK02

...


Distribution of GDP per Capita by Country
GDP per Capita
9
6
3
0
DK
FR
HU
PT
SK
Country

... p. 116

4.2 Discussion

5 Assignment 2: Cross-sectional Estimates

5.1 2A: Growth and Inequity

5.1.1 1. Model Estimation

$GINI = \beta_0 + \beta_1 \cdot \text{GDP per capita} + \epsilon$

Gini is the dependent variable (y), while GDP per capita is the independent variable (x).

	All
(Intercept)	0.055 ***
	(0.010)
log_GDP_per_capita	0.019
	(0.011)
r.squared	0.080
adj.r.squared	0.055
statistic	3.142
p.value	0.085

*** p < 0.001; ** p < 0.01; * p < 0.05.

Regression statistics of all countries for the year 2010

Beta 0 = 0.0551 (intercept)

Beta 1 = 0.0193 (slope coefficient)

Regression statistics of all countries seperately for the year 2010

5.1.2 2. Model Diagnostics

We'll now look at some of the numbers we got from the linear regression model (for all countries combined):

- Coefficients:
  - Intercept is 0.0551 (expected value of gini when GDP per capita is 0). Statistically significant (indicated by p-value).

	SK	DK	HU	PT	FR
(Intercept)	0.069	-0.127	0.075 **	0.079	-0.028
	(0.010)	(0.085)	(0.010)	(0.015)	(0.026)
log_GDP_per_capita	-0.018	0.124	0.049	-0.008	0.107 ***
	(0.034)	(0.059)	(0.029)	(0.030)	(0.026)
r.squared	0.221	0.689	0.410	0.067	0.457
adj.r.squared	-0.558	0.534	0.262	-0.866	0.430
statistic	0.284	4.432	2.778	0.072	16.820
p.value	0.689	0.170	0.171	0.833	0.001

*** p < 0.001; ** p < 0.01; * p < 0.05.

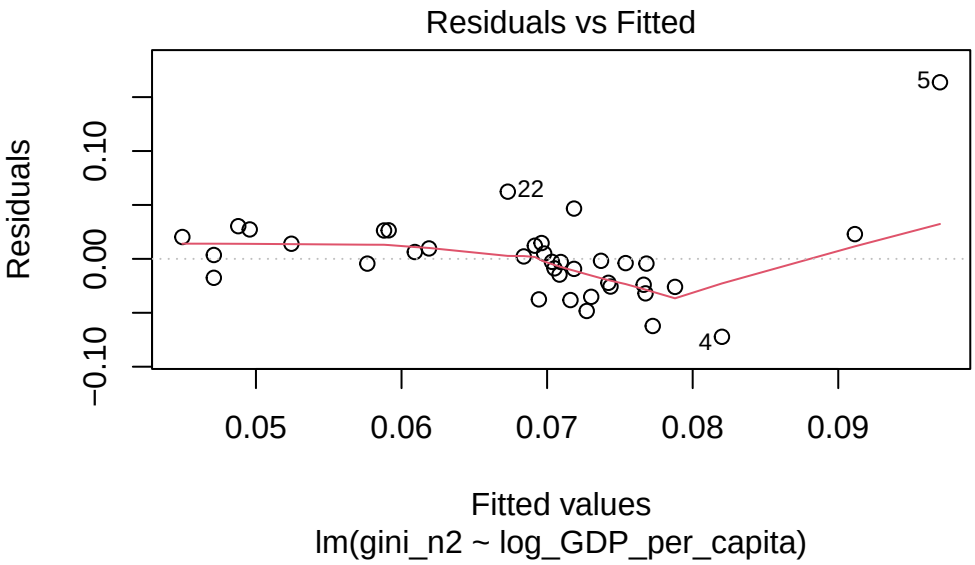
  - Estimated coefficient for GDP per capita is 0.0193, if the natural logarithm of GDP per capita increase by one, then the gini coefficient will increase by 0.0193. The p-value associated with the coefficent is however not statistically significant at 5% level.

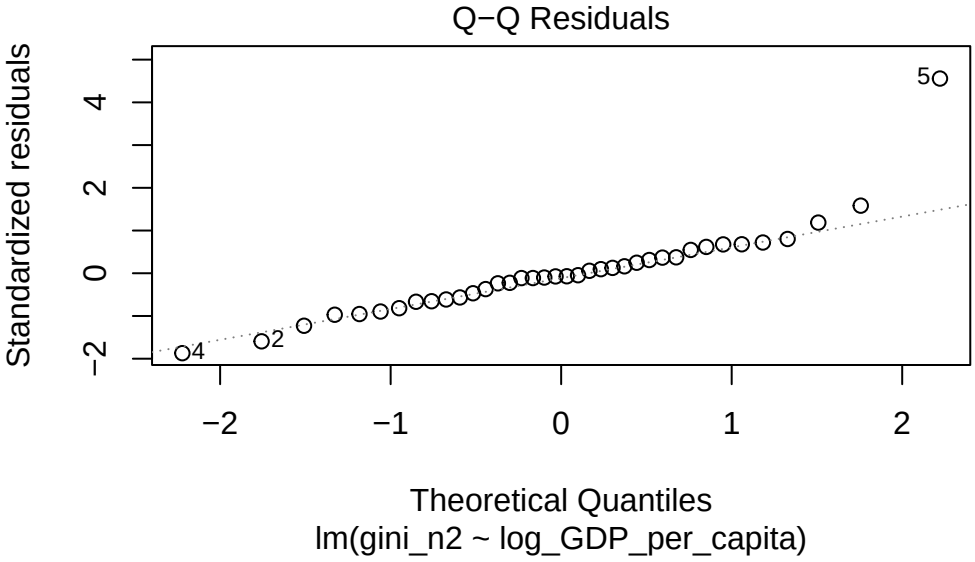
- Goodnes of fit:
  - Multiple R-squared (0.08028) indicate that around 8% of the variability in the GINI coefficient is explained by the model. This is low, which can suggest that the model dosen't explain the variation in gini.
  - Adjusted R-squared is even lower (0.05474), is therefore expected that the model dosen't really explain the variance in the dependent variable (gini).
- Model significance:
  - The F-statistic (3.142) indicate the significance of the regression model. We can see here, that with the p-value of 0.08474, the model isn't significant at a 5% level.

By looking at these numbers, we can see that the model may not reliably predict the gini coefficient. It also suggest that GDP per capita may not be a valid predictor of the gini coefficient.

We also examined our selected countries seperately in order to see how the reliability and validity of the model might vary between countries. However, since there are too few observations for most of the countries, it makes it hard to make a conclusion of the reliability. What we can see from this examination, is that the models for Denmark, Portugal and Slovakia are not statistically significant. France seem however to have significant coefficiants, and Hungary have a moderate R-squared (but lacks significance in the slope).

5.1.3 3. Visualization





We've made two plots that can help us understand the relationship between GDP per capita and the Gini coefficient. The first one is a plot that shows us residuals vs fitted values, which can help us check the homoscedasticity assumption of a linear regression model. The residuals should be randomly scattered around the horisontal 0 line, something that can inducate that the variances of the error terms are constant. In our plot, the residuals are

in some extent randomly distributed, and there is also no clear pattern; this suggest that there is likely no significant issues with heteroscedasticity or non-linearity.

The other plot - normal Q-Q plot - is used to assess if the residuals of the linear model are normally distributed. In our plot, most of the points follow the line closely, suggesting that the residuals are normally distributed.

In both of these plots, we can see that there are some outliers, these can affect the fit of the model. These outliers might come from regions that have a high GDP per capita compared to rest of the regions, like Paris in France that is a financial hub.

The numbers we gathered earlier, as well as these plots, can help us explain and assert whether the classical OLS assumptions holds for our model:

- Linearity:
  –
- Independence:
  –
- Normality:
  –
- Homoscedasticity:
  –
- Multicollinearity:
  –

```
`geom_smooth()` using formula = 'y ~ x'
```

In this plot we are visualizing the relationship between GDP per capita and gini by using a mix of the two previous plots. We can here as well, see extreme outliers.

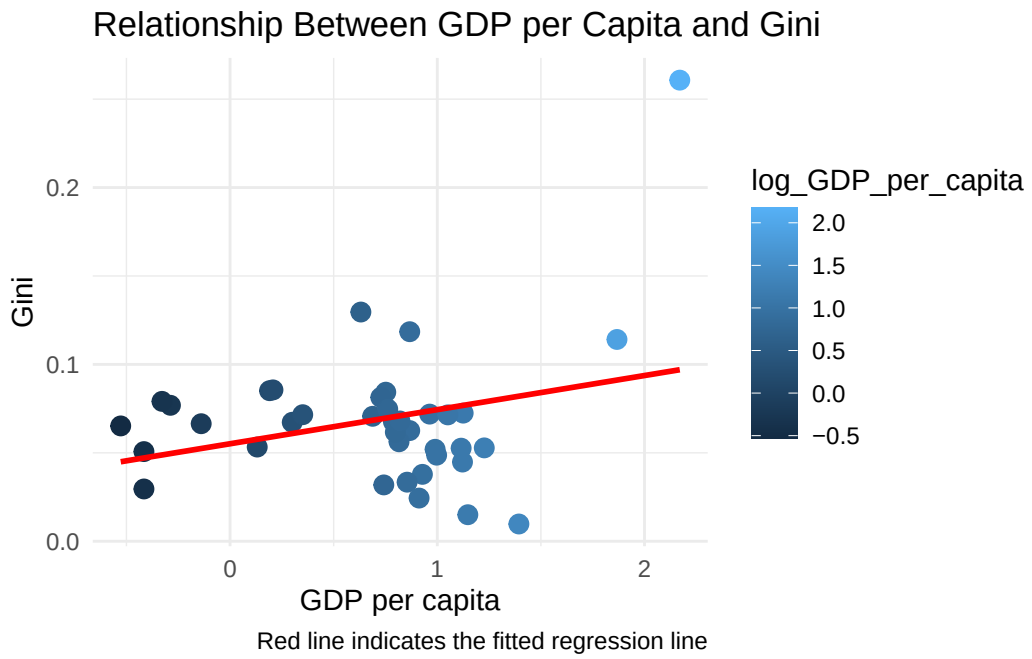


Figure 1: Relationship Between GDP per Capita and Gini

5.2 2B: Exploring Other Determinants of Inequity

5.2.1 1. Data Acquistion

In order to conduct a Multiple Linear Regression model, we need to have some independent variables to use in the model and compare them with the dependent variable. The first variable, education, can explain income inequality, since it can influence income distribution in a region. If access to education is unequal, then higher education levels might incease income disparities (Rodriguez-Pose & Tselios, 2008). The population density, our second variable, can explain inequality since regions with a higher population density might have different economic behaviours. Our last variable, rail network (infrastructure), can influence economic development and accessibility, which also can affect income inequality in a region (Chatterjee & Turnovsky, 2012).

5.2.2 2. Multiple Linear Regression Model

We will in this part do a Multiple Linear Regression model by using the variables education (in percentage of pupils and students in education, % of total population), population density and rail network in km. This model will tell us if these variables can help explain change in the gini coefficient.

In both our simple linear regression model, and now in our multiple linear regression model, we use the logarithmic function which makes it easier to linearize the relationship between the variables. By using it for GDP per capita, we can reflect changes more effectively. For

rail network and population density, the logarithm function ensure that the model capture proportional changes and deals better with the wide range of values.

	Model	Model 2	Model 3
(Intercept)	0.050	0.008	-0.206
	(0.091)	(0.078)	(0.114)
log_GDP_per_capita	0.020	0.011	0.017
	(0.014)	(0.012)	(0.012)
students_percentage	0.000	-0.006	-0.001
	(0.004)	(0.004)	(0.006)
log(pop_density)		0.041 **	0.032 *
		(0.012)	(0.014)
log(rail_km)			0.020
			(0.018)
r.squared	0.098	0.370	0.467
adj.r.squared	0.033	0.300	0.370
statistic	1.515	5.284	4.812
p.value	0.237	0.005	0.006

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Multiple Linear Regression Model

5.2.3 3. Model Interpretation

Our first model in the Multiple Linear Regression model examine how education in addition to GDP per capita can help explain the gini coefficient. The second model look at both education, GDP per capita, and population density, while the third model examine them all and also add rail network.

Conclusion

Chatterjee, S., & Turnovsky, S. J. (2012). Infrastructure and inequality. *European Economic Review*, *56*(8), 1730–1745.

Eurostat. (2021). *Population change - Demographic balance and crude rates at regional level (NUTS 3) (demo_r_gind3)*. https://ec.europa.eu/eurostat/cache/metadata/en/demo_r_gind3_esms.ht

Eurostat. (2023a). *Gross domestic product (GDP) at current market prices by NUTS 3 regions*. https://ec.europa.eu/eurostat/databrowser/view/nama_10r_3gdp/default/table?lang=en.

Eurostat. (2023b). *Population on 1 January by broad age group, sex and NUTS 3 region*. https://ec.europa.eu/eurostat/databrowser/view/demo_r_pjanaggr3/default/table?lang=en.

Eurostat. (2023c). *Regional economic accounts (reg_eco10)*. https://ec.europa.eu/eurostat/cache/metadata/

Hasell, J., & Roser, M. (2023). Measuring inequality: What is the Gini coefficient? *Our World in Data*.

Rodriguez-Pose, A., & Tselios, V. (2008). *EDUCATION AND INCOME INEQUALITY IN THE REGIONS OF THE EUROPEAN UNION*. https://click.endnote.com/viewer?doi=10.1111%2Fj.1467-9787.2008.00602.x&token=WzM0MzQ0MzksIjEwLjExMTEvai4xNDY3LTk3ODcuMjAwOC4wMDYwMi5Avuk84iH8myPb7qaaaJIQk.

Wooldridge, J. M. (2020). *Introductory Econometrics - A Modern Approach* (7th ed.). Cengage.