Eurostat EDA

Sindre H. Øveraas, Alen Colacovic & Sebastian M. Fløysand

Eurostat EDA

In this paper we are going to present an explanatory data analysis of different statistics of selected European countries. We are more specifically going to analyze the European countries Austria (AT), Belgium (BE), Bulgaria (BG), Croatia (HR), Italy (IT), Serbia (RS), and Sweden (SE).

The paper will consist of four parts (assignments), where in the first part we will explore sub-national GDP and regional inequity in our selected countries. For this part we will use data collected from Eurostat. The data for the first part of the paper consists of GDP and population statistics for the years 2000 – 2020, on a sub-regional level i.e., at NUTS 3 level. NUTS (Nomenclature of territorial units for statistics) is the geographical nomenclature subdividing the economic territory of countries in the European Union. These levels consist of NUTS 1, 2 and 3, with 3 representing the smallest territorial units in a country (*Glossary*, 2021). The remaining parts of this paper i.e., 2, 3 and 4, will be explained continuously and gradually when we eventually get to them later in the paper.

To start our analysis of sub-national GDP and regional inequity for our selected countries we must, as mentioned, collect data from Eurostat. We download population by broad age group and sex, as well as gross domestic product at current markets prices, at NUTS 3 level (*Population on 1 January by Broad Age Group, Sex and NUTS 3 Region*, 2022) (*Gross Domestic Product (GDP) at Current Market Prices by NUTS 3 Regions*, 2022). After we have added our two datasets to our RStudio project, we can calculate GDP per capita at the NUTS 3 level for the separate countries. This is achieved with dividing the GDP on the number of population figures, and can be presented with the following formula:

$$y_i = GDP_i/population_i$$

Sub-National GDP

	GDP	Population	GDP_capita
	Min. : 74.55	Min. : 20320	Min. : 1087
	1st Qu.: 1738.28	1st Qu.: 164518	1st Qu.:17180
	Median : 5614.05	Median : 273920	Median :25185
	Mean : 10238.24	Mean : 406217	Mean :24191
	3rd Qu.: 10640.23	3rd Qu.: 429030	3rd Qu.:31351
	Max. :181212.88	Max. :4355725	Max. :72062
	NA	NA's :771	NA's :771

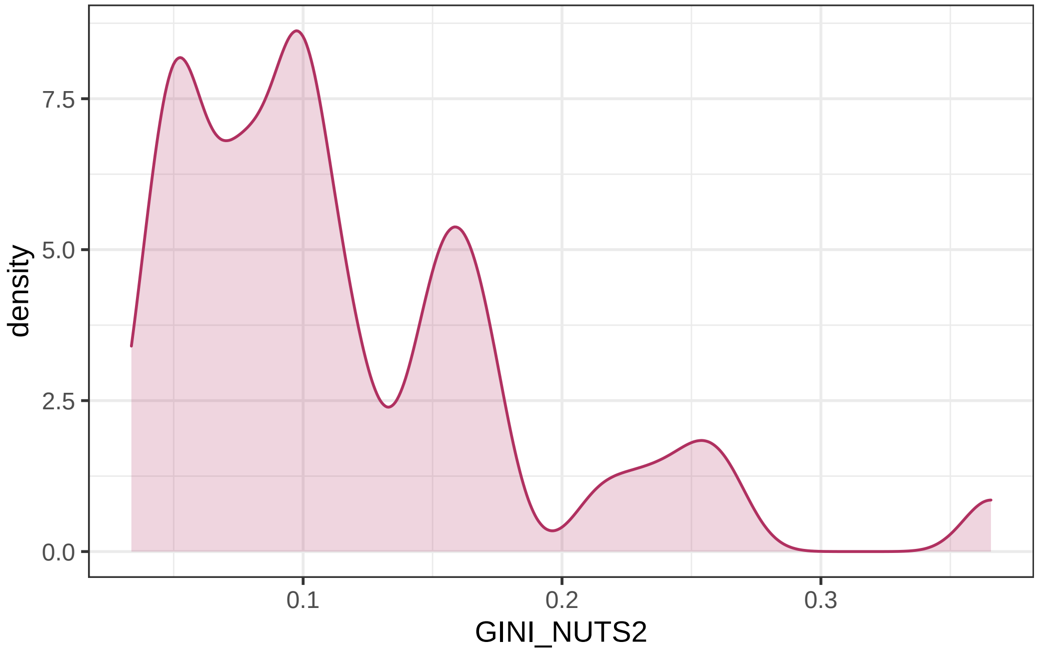
Looking at the GDP per capita for all countries in the dataset, we find that the difference in population from min to max is big. The difference between Median and Mean is also relatively big witch can indicate that some of the biggest regions has much larger amount of population then the rest and therefore affects the mean and pulls it higher. The fact that the 3rd quartile is just a few thousands away from the mean amplifies our suspicion that we have some outliers with a very high population compared to the rest.

Its reason to believe that high population equals high GDP based on the fact that it is more people that contributes to the GDP. However, this can not be applied in every circumstances. For example Monaco with a population in 2020 of just above 39.000 (**MonacoPopulation2022?**) had a GDP on 6.25 billion USD the same year versus Burundi with a GDP on 3.22 billion (**GDPConstant2015?**) and a population just above 11.89 million (**BurundiPopulation2022?**). The GDP per capita gives us a more accurate measure. in the model above we can see that its a big difference in GDP per capita in each region, this can be caused by a population or a cluster with rich/poor individuals.

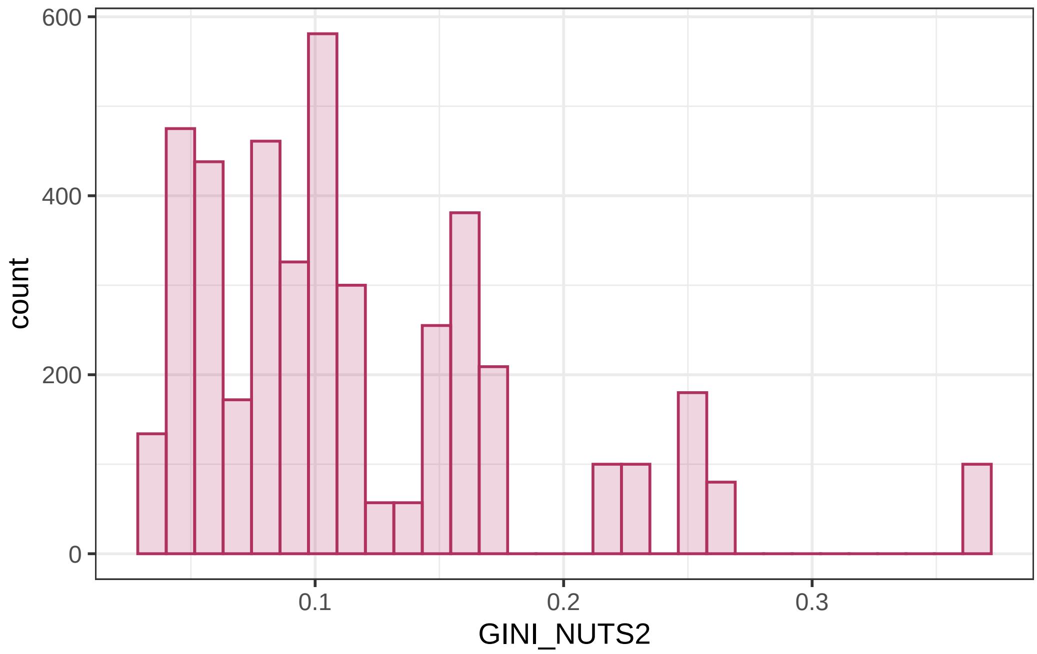
We also had some NA values witch we chose to take out of the dataset. NA values may come from the fact that Population or the GDP was not measured this year or was not available when the dataset was made.

```
[1] 0.2603924
```

	GINI_NUTS2
	Min. :0.03367
	1st Qu.:0.07065
	Median :0.09839
	Mean :0.11800
	3rd Qu.:0.15525
	Max. :0.36569

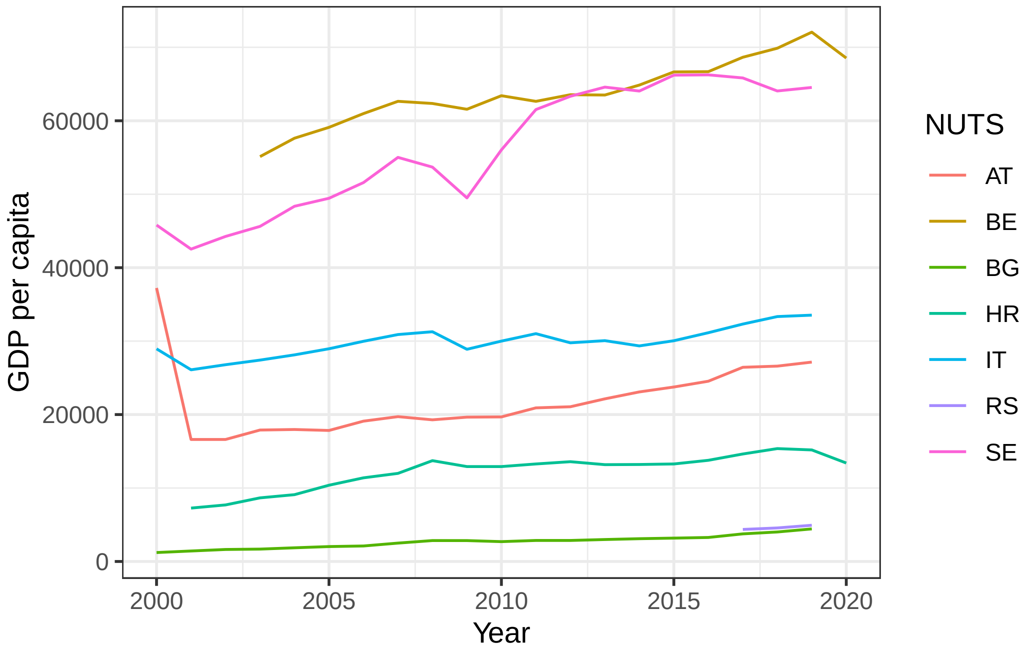



```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



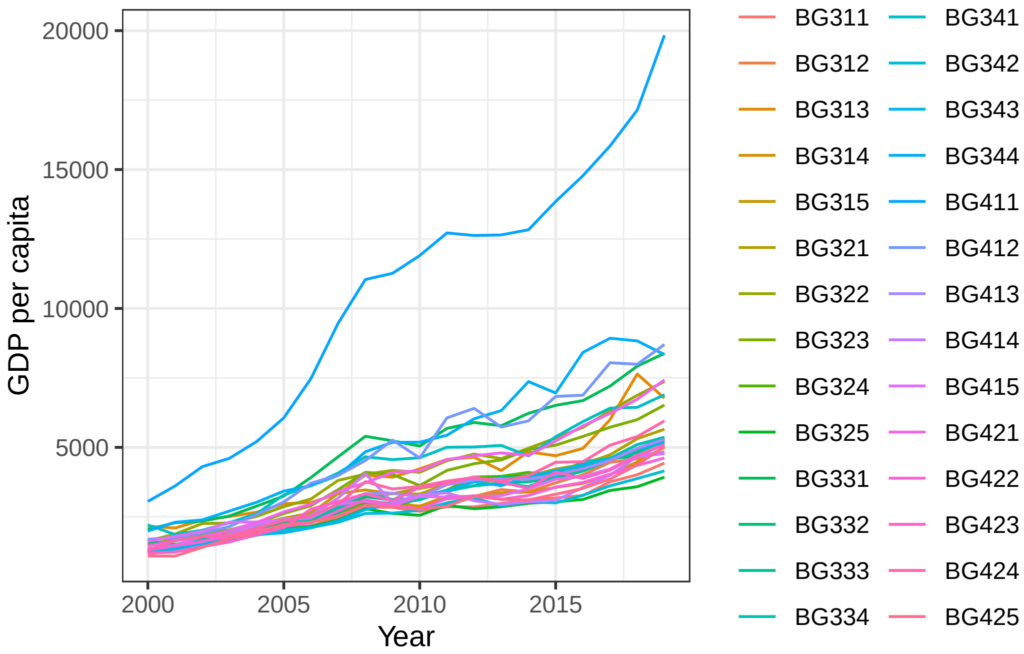
Looking at the plot above there is one outlier up against 0.4 with around 100 observations. The same result also seem to accure in the density plot.

```
GDP_Per_Capita %>% distinct(NUTS, Year, .keep_all = TRUE) %>%
  ggplot(aes(x = Year, y = GDP_capita, colour = NUTS)) + geom_line(lwd = .5) + labs(x =
```



Bulgaria GDP

```
GDP_Per_Capita_BG <- GDP_Per_Capita %>%
  filter(NUTS == "BG" ) %>% select(GDP_capita, Region) %>%
  select(Region)

GDP_Per_Capita %>%
  filter(Region %in% GDP_Per_Capita_BG$Region) %>%
  ggplot(aes(x = Year, y = GDP_capita, colour = Region)) + geom_line(lwd = .5) + labs(x
```




```
GDP_Per_Capita %>%
  filter(NUTS == "BG", Year == 2010) %>%
  select(Region, GDP_capita) %>%
  slice_max(GDP_capita, n = 3)
```

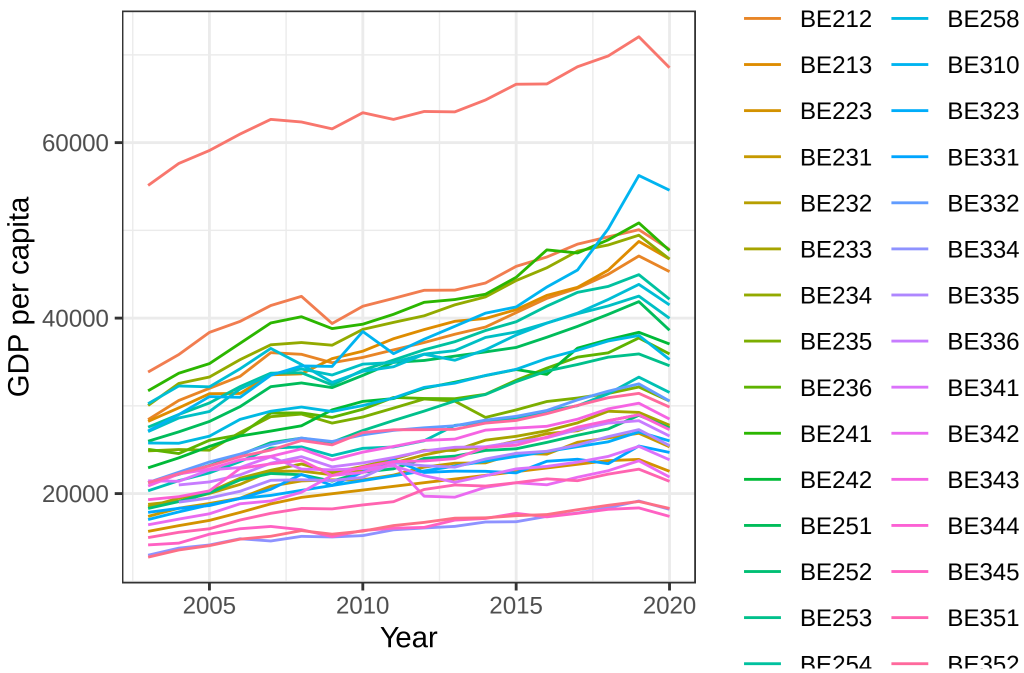
```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 BG411      11905.
2 BG344       5187.
3 BG331       5049.
```

```
GDP_Per_Capita %>%
  filter(NUTS == "BG", Year == 2010) %>%
  select(Region, GDP_capita) %>%
  slice_min(GDP_capita, n = 3)
```

```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 BG325       2555.
2 BG311       2701.
3 BG342       2735.
```

Belgium GDP

```
GDP_Per_Capita_BE <- GDP_Per_Capita %>%
  filter(NUTS == "BE" ) %>% select(GDP_capita, Region) %>%
  select(Region)

GDP_Per_Capita %>%
  filter(Region %in% GDP_Per_Capita_BE$Region) %>%
  ggplot(aes(x = Year, y = GDP_capita, colour = Region)) + geom_line(lwd = .5) + labs(x
```




```
GDP_Per_Capita %>%
  filter(NUTS == "BE", Year == 2010) %>%
  select(Region, GDP_capita) %>%
  slice_max(GDP_capita, n = 3)
```

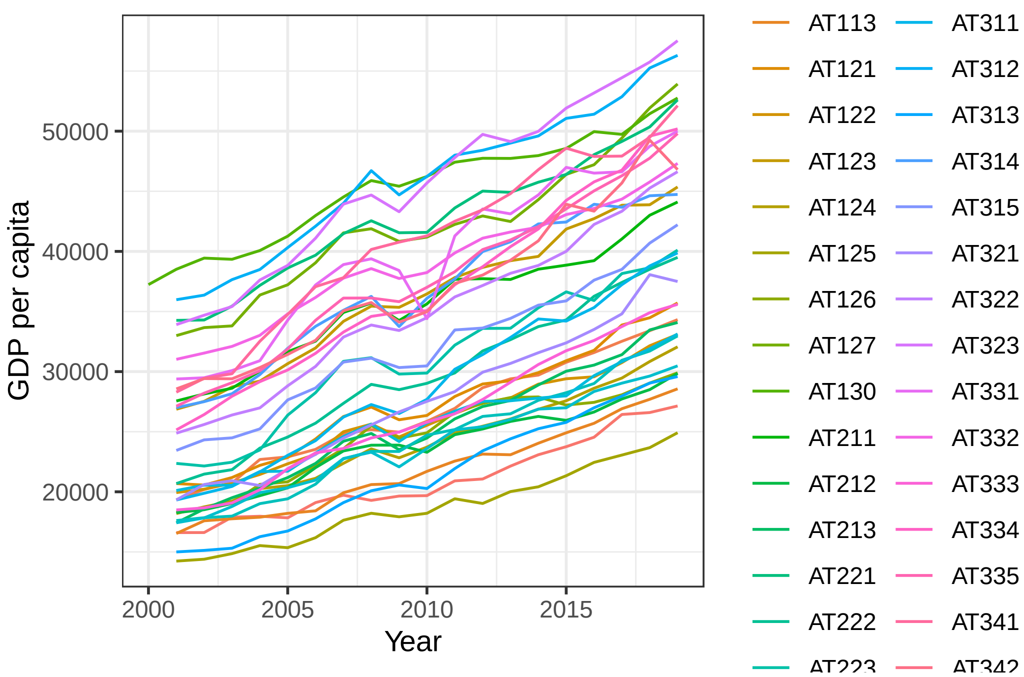
```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 BE100      63409.
2 BE211      41353.
3 BE241      39307.
```

```
GDP_Per_Capita %>%
  filter(NUTS == "BE", Year == 2010) %>%
  select(Region, GDP_capita) %>%
  slice_min(GDP_capita, n = 3)
```

```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 BE334      15212.
2 BE353      15727.
3 BE345      15797.
```

Austria GDP

```
GDP_Per_Capita_AT <- GDP_Per_Capita %>%
  filter(NUTS == "AT" ) %>% select(GDP_capita, Region) %>%
  select(Region)

GDP_Per_Capita %>%
  filter(Region %in% GDP_Per_Capita_AT$Region) %>%
  ggplot(aes(x = Year, y = GDP_capita, colour = Region)) + geom_line(lwd = .5) + labs(x
```



```r
GDP_Per_Capita %>%
  filter(NUTS == "AT") %>%
  select(Region, GDP_capita) %>%
  slice_max(GDP_capita, n = 3)
```

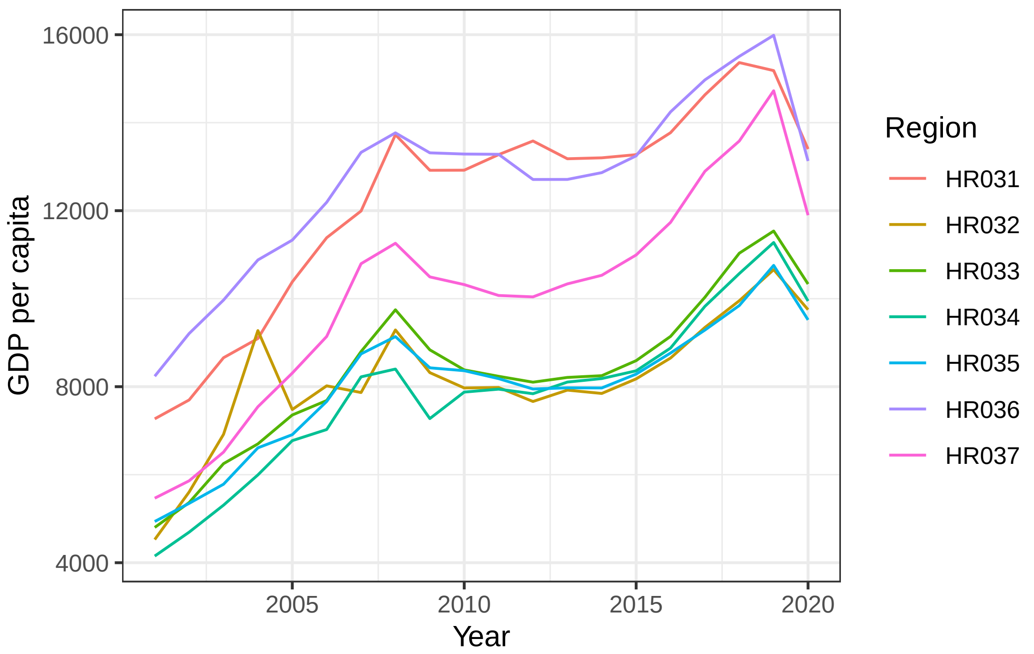
```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 AT323      57525.
2 AT312      56307.
3 AT323      55748.
```

```r
GDP_Per_Capita %>%
  filter(NUTS == "AT") %>%
  select(Region, GDP_capita) %>%
  slice_min(GDP_capita, n = 3)
```

```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 AT125      14236.
2 AT125      14392.
3 AT125      14858.
```

Croatia GDP

```r
GDP_Per_Capita_HR <- GDP_Per_Capita %>%
  filter(NUTS == "HR" ) %>% select(GDP_capita, Region) %>%
  select(Region)

GDP_Per_Capita %>%
  filter(Region %in% GDP_Per_Capita_HR$Region) %>%
  ggplot(aes(x = Year, y = GDP_capita, colour = Region)) + geom_line(lwd = .5) + labs(x
```




```r
GDP_Per_Capita %>%
  filter(NUTS == "HR") %>%
  select(Region, GDP_capita) %>%
  slice_max(GDP_capita, n = 3)
```

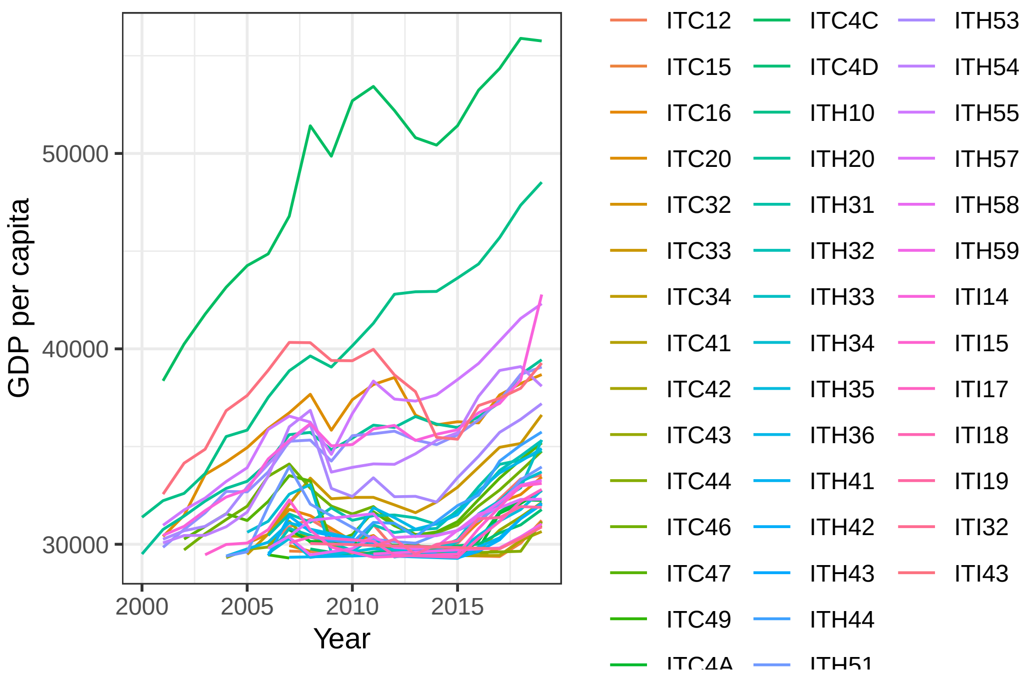
```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 HR036      15986.
2 HR036      15507.
3 HR031      15366.
```

```r
GDP_Per_Capita %>%
  filter(NUTS == "HR") %>%
  select(Region, GDP_capita) %>%
  slice_min(GDP_capita, n = 3)
```

```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 HR034       4151.
2 HR032       4526.
3 HR034       4694.
```

Italy GDP

```
GDP_Per_Capita_IT <- GDP_Per_Capita %>%
  filter(NUTS == "IT" ) %>% select(GDP_capita, Region) %>%
  select(Region)

GDP_Per_Capita %>%
  filter(Region %in% GDP_Per_Capita_IT$Region) %>%
  slice_max(GDP_capita, n = 500) %>%
  ggplot(aes(x = Year, y = GDP_capita, colour = Region)) + geom_line(lwd = .5) + labs(x
```




```
GDP_Per_Capita %>%
  filter(NUTS == "IT") %>%
  select(Region, GDP_capita) %>%
  slice_max(GDP_capita, n = 3)
```

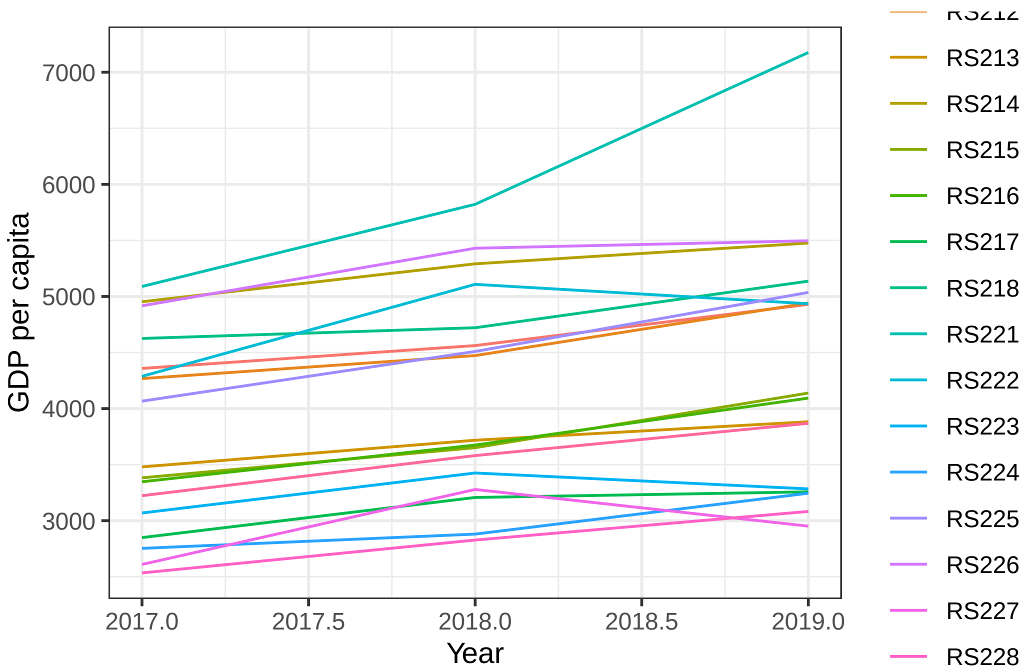
```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 ITC4C      55890.
2 ITC4C      55756.
3 ITC4C      54347.
```

```
GDP_Per_Capita %>%
  filter(NUTS == "IT") %>%
  select(Region, GDP_capita) %>%
  slice_min(GDP_capita, n = 3)
```

```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 ITF64      11713.
2 ITG14      11977.
3 ITG16      12311.
```

Serbia GDP

```
GDP_Per_Capita_RS <- GDP_Per_Capita %>%
  filter(NUTS == "RS" ) %>% select(GDP_capita, Region) %>%
  select(Region)

GDP_Per_Capita %>%
  filter(Region %in% GDP_Per_Capita_RS$Region) %>%
  ggplot(aes(x = Year, y = GDP_capita, colour = Region)) + geom_line(lwd = .5) + labs(x
```



```r
GDP_Per_Capita %>%
  filter(NUTS == "RS") %>%
  select(Region, GDP_capita) %>%
  slice_max(GDP_capita, n = 3)
```

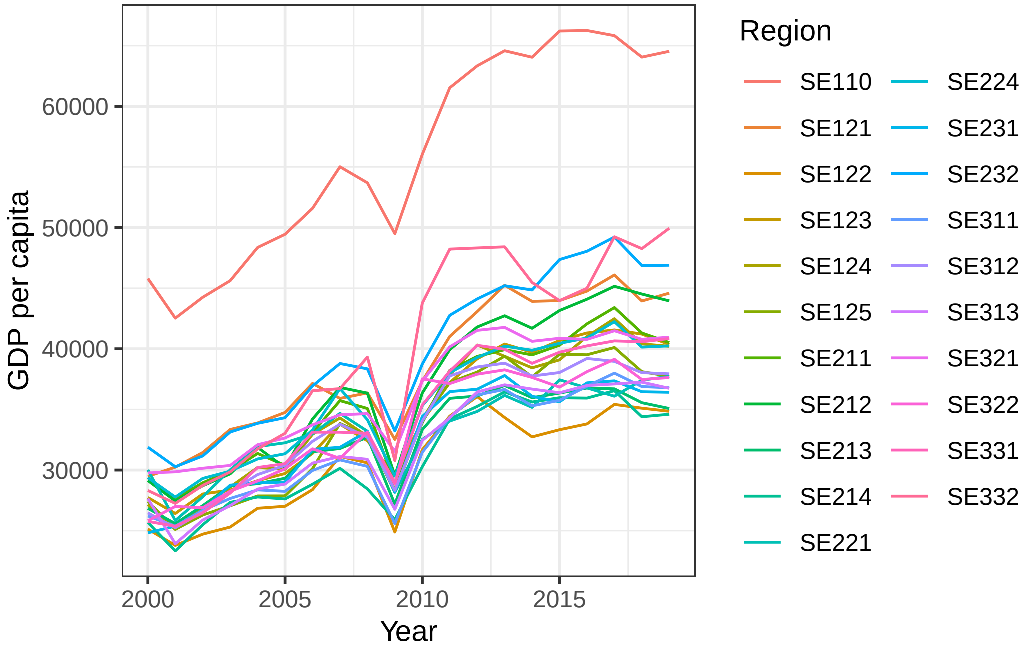
```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 RS221       7176.
2 RS221       5822.
3 RS226       5497.
```

```r
GDP_Per_Capita %>%
  filter(NUTS == "RS") %>%
  select(Region, GDP_capita) %>%
  slice_min(GDP_capita, n = 3)
```

```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 RS228       2534.
2 RS227       2610.
3 RS224       2753.
```

Sweden GDP

```r
GDP_Per_Capita_SE <- GDP_Per_Capita %>%
  filter(NUTS == "SE" ) %>% select(GDP_capita, Region) %>%
  select(Region)

GDP_Per_Capita %>%
  filter(Region %in% GDP_Per_Capita_SE$Region) %>%
  ggplot(aes(x = Year, y = GDP_capita, colour = Region)) + geom_line(lwd = .5) + labs(x
```




```
GDP_Per_Capita %>%
  filter(NUTS == "SE") %>%
  select(Region, GDP_capita) %>%
  slice_max(GDP_capita, n = 3)
```

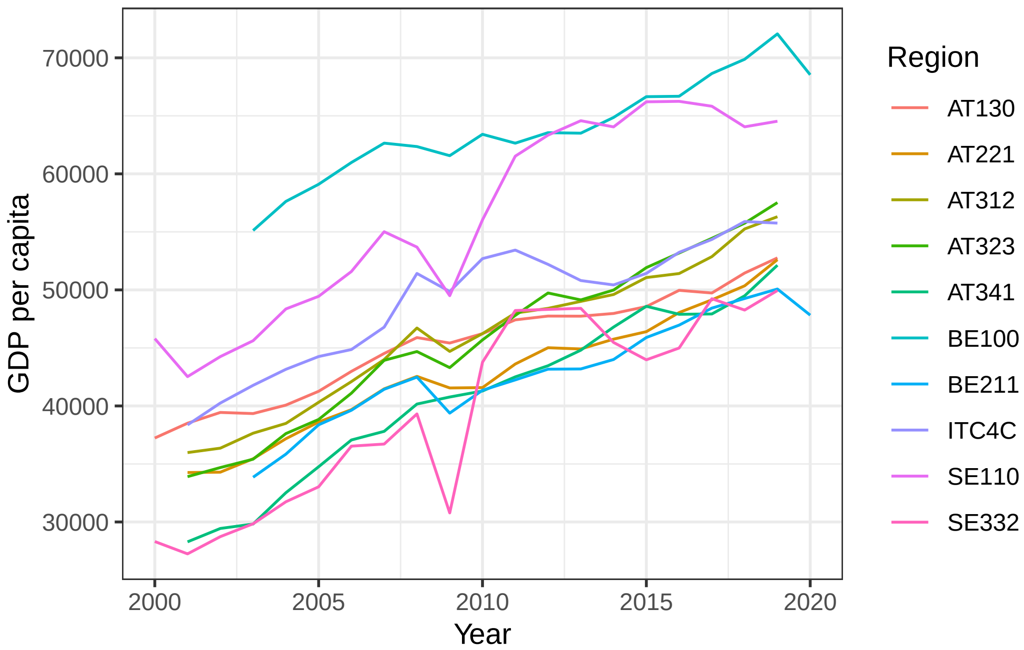
```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 SE110      66250.
2 SE110      66209.
3 SE110      65827.
```

```
GDP_Per_Capita %>%
  filter(NUTS == "SE") %>%
  select(Region, GDP_capita) %>%
  slice_min(GDP_capita, n = 3)
```

```
# A tibble: 3 x 2
  Region GDP_capita
  <chr>       <dbl>
1 SE214      23331.
2 SE122      23782.
3 SE313      23921.
```

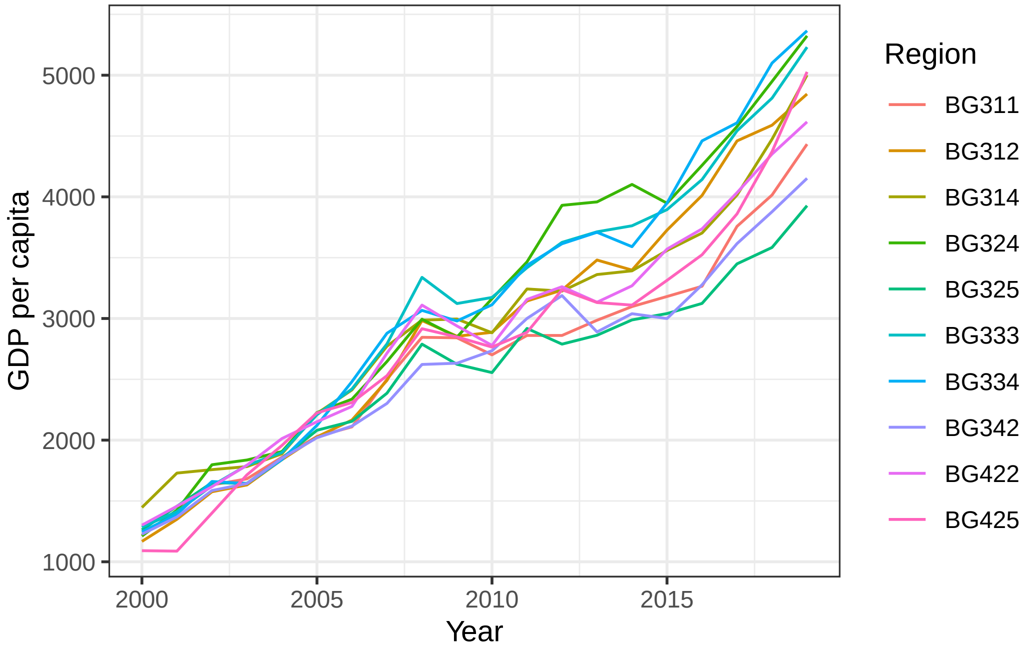
10 regions with highest GDP

```
GDP_Per_Capita_Max <- GDP_Per_Capita %>%
  filter(Year == 2010) %>% select(GDP_capita, Region) %>%
  slice_max(GDP_capita, n = 10) %>%
  select(Region)

GDP_Per_Capita %>%
  filter(Region %in% GDP_Per_Capita_Max$Region) %>%
  ggplot(aes(x = Year, y = GDP_capita, colour = Region)) + geom_line(lwd = .5) + labs(x
```



10 regions with lowest GDP

```
GDP_Per_Capita_Min <- GDP_Per_Capita %>%
  filter(Year == 2010) %>% select(GDP_capita, Region) %>%
  slice_min(GDP_capita, n = 10) %>%
  select(Region)

GDP_Per_Capita %>%
  filter(Region %in% GDP_Per_Capita_Min$Region) %>%
  ggplot(aes(x = Year, y = GDP_capita, colour = Region)) + geom_line(lwd = .5) + labs(x
```



Glossary:Nomenclature of territorial units for statistics (NUTS). (2021). https://ec.europa.eu/eurostat/statis explained/index.php?title=Glossary:Nomenclature_of_territorial_units_for_statistics_(NUTS).

Statistics | Eurostat. (2022). https://ec.europa.eu/eurostat/databrowser/view/nama_10r_3gdp/default/tab

Statistics | Eurostat. (2022). https://ec.europa.eu/eurostat/databrowser/view/demo_r_pjanaggr3/default/t