**Goal**

The main goal of the Data Visualization final project is to tell a coherent story from a complex agro-environmental dataset.

**Work steps**

1. Questions/hypothesis definition
2. Preparation of datasets, through database queries and data wrangling
3. Summary statistics
4. Exploratory data analysis
5. Inferential statistics
6. Final visualisation product and storytelling

**Assessment**

1. ~~A~~ **~~live presentation~~** ~~of your story (it may consist on the planned work with a preliminary data exploratory analysis/description).~~
2. A short **written report**, including the code as an Appendix.

**Report**

Should include the following chapters:

1. **Introduction** – Short introduction to the topic, ending with questions/working hypothesis to be addressed and objective (2 pages max.)
2. **Database description** – Short descriptive statistics of the database/tables (2 page max.)
3. **Exploratory data analysis** – This will be the most important chapter, where you will try to tell a coherent history by means of numerical outputs and visualizations (10 pages max.)
4. **Discussion/Conclusions** – a short discussion/main take home messages/conclusions of the work (2 pages max.).
5. **References**
6. **ANNEX – Python code.**

**6. Analysis of agricultural activities: temporal trends**

Objectives: (1) Explore how agriculture practices have changed over time; (2) How these changes of practices are related with changes in socio-economic indicators?

Data available: The available information are numbers per civil parishes. Proportions should be used to allow comparisons over time.

Types of analysis (examples): ordination techniques with temporal dimension (for example defining temporal pathways within the multivariate space); time series analysis.

Visualizations (examples): Line plots describing temporal trends; multivariate analysis outputs (e.g. PCA biplots); possibly Sankey/alluvial diagrams may also be appropriate to describe changes over time.

Contents

[Introduction 2](#_Toc169275813)

[Data 2](#_Toc169275814)

[Data Analyses 4](#_Toc169275815)

[Discussion 4](#_Toc169275816)

[Conclusion 4](#_Toc169275817)

[References 5](#_Toc169275818)

[ANNEX – Python code 5](#_Toc169275819)

# Introduction

In this project we aim to explore **how agriculture practices have changed over time; and see how these changes of practices are related with changes in socio-economic indicators.** The source of data for this project comes from the INE database, a Portuguese database containing a volume of Portuguese census information. For our purposes we were interested in gathering information on the agricultural activities and socio-economic indicators. The INE database contains agricultural activities involving permanent, temporary, grassland crops and livestock grouped by ‘freguesia’ or town with data from the years 1989, 1999, 2009, and 2019. Each of these topics also included information on the the … (refer to table 1).

Futhermore, we wanted socio-economic indicators which we decided would be production values, education levels, and labor types and counts also grouped by each town in the 1989, 1999, 2009, and 2019 (refer to table 2). With information about both the agricultural activities happening in each town over the course of 40 years and information on how socio-economic conditions have changed over 40 years, we could demonstrate a lot about the changes that have taken place over time and make further predictions about how such practices have changed and relate to socio-economic factors. This information is both necessary and interesting to analyze given the environmental changes and socio-economic changes that have taken place in the last 40 years. The world is only becoming more fast-paced and more extreme environmental changes are taking place, because of this, farmers will need to adapt their practices. Conversely, at the beginning of this project we were under the assumption that socio-economic conditions have improved in our societies. Therefore this project acts as an interesting way to verify this assumption and look into what improved “socio-economic” conditions actually imply.

~~Specifically, the hypotheses we decided to test are:~~

1. ~~The variety of agricultural practices have increased over time in each region.~~
2. ~~The larger variety of agricultural practices, the more economic gain (in production\_euro) and the greater the labour force.~~
3. ~~The more greater variety in ag practices, the greater the economic gain in production\_euro~~

# Data

The data that was used for this work was extracted from the excel spreadsheet of INE database provided by professor Pedro Segurado. In this table we aim to use columns like \*\*\*\*

1. **Database description** – Short descriptive statistics of the database/tables (2 page max.)

* **Discuss the INE database more**?
* **Discuss how the database was normalized in SQL and then uploaded as a csv file**

Table 1: Agricultural Activity Data

|  |  |
| --- | --- |
| **Agricultural Activity** | **About** |
| Agricultural holdings with permanent crops |  |
| Agricultural holdings with temporary crops (No.) | Cereals, Dried Pulses, Potatoes, Sugarbeets, Fresh vegetables, Flowers and Ornamental Plants, Industrial crops, Other temporary crops |

Table 2: Socio-Economic Indicator Data

|  |  |
| --- | --- |
| **Socio-Economic Indicator** | **About** |
| Production value (euro) | Value of total standard production of agricultural holdings (in euro) |
| labour\_regular | Volume of agricultural labour force (AWU) and Type of labour force - Regular |
| Labour\_non\_regular |  |
| Labur not hired | Volume of agricultural labour force (AWU) and Type of labour force - Workers not hired by the holder |
| Labour family |  |
| Edu\_none | Number of farmers without education by freguesia |
| Edu\_basic | Number of farmers with basic education by freguesia |
| Edu\_secondary | Number of farmers with secondary education by freguesia |
| Edu\_superior | Number of farmers with superior education by freguesia |

# Data Analyses

1. **Exploratory data analysis** – This will be the most important chapter, where you will try to tell a coherent history by means of numerical outputs and visualizations (10 pages max.)

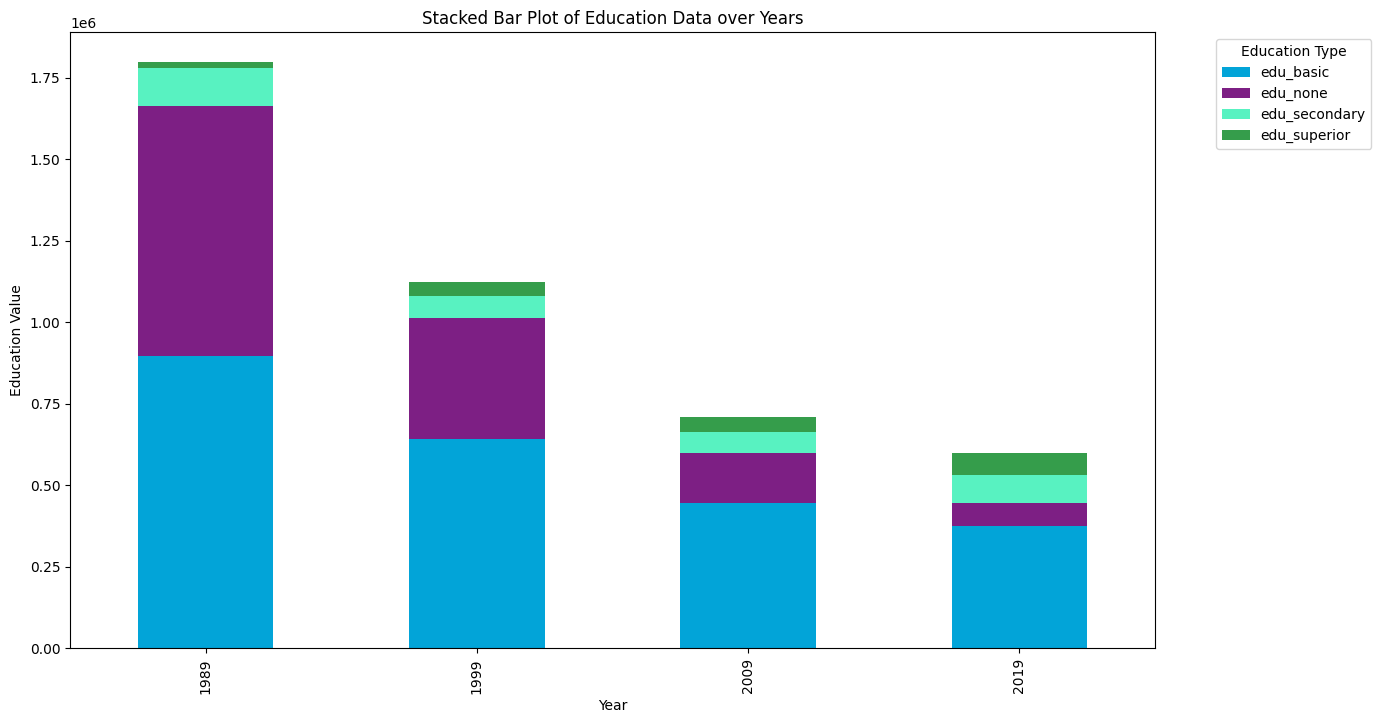
To begin our exploratory data analysis, we began by looking at how our data was distributed. The main aspects of the data that we were looking at were the temporary and permanent agricultural activities broken down by the count of holding per freguesia (county) and the socio-economic factors ---labour type, education level, and production value, in each county all divided into the years 1989, 1999, 2009, and 2019. We used histograms and scatter plots to understand the data distribution.

Here we break down the data distribution by our categories of interest:

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

This education information histogram demonstrates that a high count of farmers have their basic education and secondary education.

The next graphic explains how this information is broken down per year. 

As one would assume, farmers with no education is highest in 1989, indicated by the large amount of purple in the stacked bar plot. And there is an increase in secondary and superior education by 2009 and 2019, as indicated by the smaller presence of the blue bar and the higher portion of the green bars. The large decrease in the bar plots is also interesting as it implies that there are less individual farmers in Portugal since 1999.

A group of blue and white rectangular shapes

Description automatically generated

**Stacked bar plots to see how ag activities have changed over time**

A graph with different colored bars

Description automatically generatedA graph with different colored bars

Description automatically generated

**Line Plots to see how Agricultural Activities have changed over time**

A graph of a crop holding

Description automatically generated with medium confidence

Similar to the bar plot, the line graph helps us to understand that the number of temporary crop holdings has greatly decreased over the years. The line graph gives a clear temporal view of the changes in the data but, unlike the stacked bar plot it is harder to tell by how much exactly this crop holding counts are changing.

A graph of growing crops

Description automatically generated

Now lets look at how the holding counts have been influenced regionally over the years:

* Graphics by region by year

**(1) Explore how agriculture practices have changed over time;**

-Diferentes graficos das praticas culturais, ou seja quais as culturas a aumentar, quais as que diminuem? Graficos por ano. Maybe por zona? Avaliar as diferentes variaçoes.

We ran a Linear Discrimant Analysis to see to what degree our data can be well separated by year. The results show that the agricultural activities and socio-economic factos can in fact be pretty well separated on a temporal scale.

A chart with dots and numbers

Description automatically generated with medium confidence

A chart with many colored dots

Description automatically generated with medium confidence

The socio-economic data, especially indicates a large degree of separation between the years 1989 and 2019.

Now, considering that the data is well separted between the years 1989 and 2019, let’s try to see how the agricultural activities and socio-economic factors are related.

-PCA sobre por ex temporary crops para ver quais as produçoes que mais afetam as variaçoes? Faz sentido? No trabalho do vasco ha esta linha: a reduction in the area of temporary crops maianly connected to the decrease in cereals production.

(2) How these changes of practices are related with changes in socio-economic indicators?

We ran a multiple linear regression model to understand which variables had the greatest significance in affecting the production\_eur, so the money made by selling each holding. The results were startling.

First after running the multiple linear regression with production\_eur as our dependent variable we found that the following agricultural and environmental variables had the creates correlation coefficient values (being far from 0).

significant\_vars = ['cereals\_count', 'grasses\_count', 'fodder\_count', 'potatoes\_count', 'sugarbeets\_count', 'industrial\_crops\_count', 'fresh\_veg\_count', 'other\_temp\_crops\_count', 'fresh\_fruit\_holdings', 'citrus\_holdings', 'vineyard\_holdings', 'edu\_none', 'edu\_superior']

Next, we plotted a bar graph showing how the average number of holdings for each variable and the average education count and related it to the production\_eur value. To our surprise, the results show that in general, the production value has decreased for these crops.

A graph with red and blue bars

Description automatically generated

-analisar as variaçoes ao longo dos anos com diferentes gráficos

🡪 plot showing how edu type and temporary crop holdings are related over time

🡪 plot shoing how labour type and temp crop holdings are related over time

🡪 plot showing how production value eur and temporary crops have changed over time

-Maybe PCA para as diferentes culturas? Não sei se faz sentido para ver quais os fatores que mais afetam a produçao de determinada cultura

# Discussion

1. **Discussion/Conclusions** – a short discussion/main take home messages/conclusions of the work (2 pages max.).

Important point to discuss, diminuição de areas não implica diminuiçao de produtividade ou de produção ao nivel do pais. Era interessante entender se estas trends vao de acordo com a produtividade do pais de cada cultura?

Será que podemos concluir alguma relação entre o aumento/diminuição de algumas culturas com factores socio economicos?

# Conclusion

# References

# ANNEX – Python code