FInal project Data Memo

install.packages('tinytex') tinytex::install_tinytex()

Final Project Data Memo

Shogo Moridaira

To make sure that each student is progressing in the final project for this course, and to verify that your data set(s) of choice will work for a machine learning project, everyone is asked to submit a check-in or "data memo."

For this assignment, you should submit **two files**

- an **.Rmd** (R Markdown) file containing your data memo

and

- the resulting knitted **.html** or **.pdf** file

Your data memo should be written and formatted neatly.

Your final project is intended to showcase your analytical abilities and model-building skills on a dataset of your choosing. Your data memo should be written in the form of a short paper; if you've already read your data file into R, you can include code or figures, but you are not required to.

Your memo should answer all of the following:

An overview of your dataset

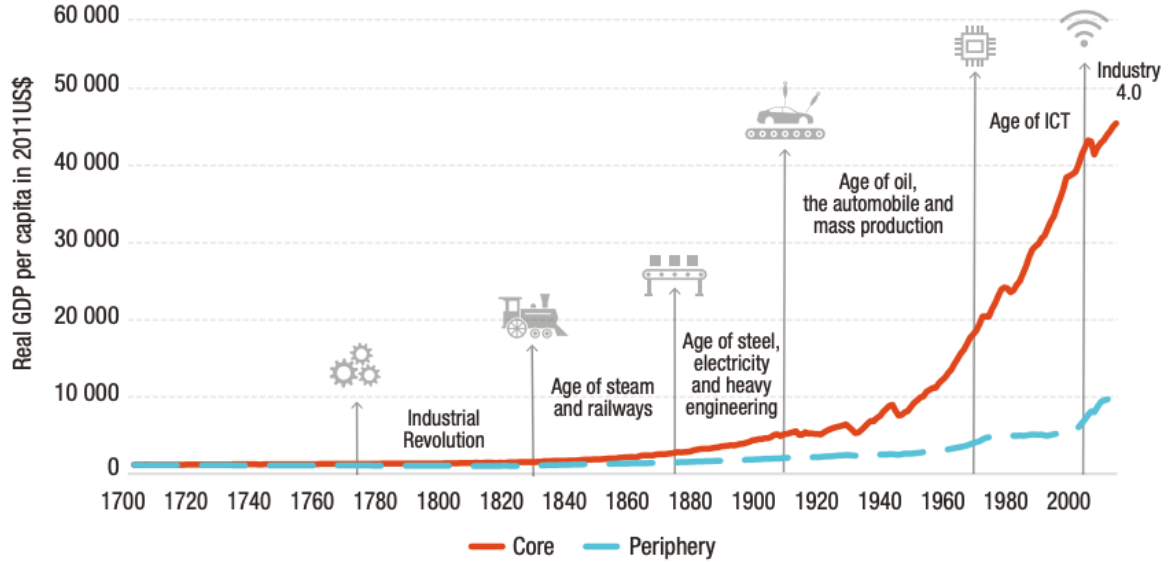
What does it include?

Where and how will you be obtaining it? Include the link and source.

- R&D Expenditure: World Bank,https://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS
- Patent Counts: World Intellectual Property Organization (WIPO)
- Technology Indices: https://unctad.org/system/files/official-document/tir2020_en.pdf
  - from ****Maddison Project Database 2018****

Figure 1

Technological change and inequality through the ages



Source: UNCTAD, based on data from Maddison Project Database, version 2018, Bolt et al. (2018), Perez (2002), and Schwab (2013).

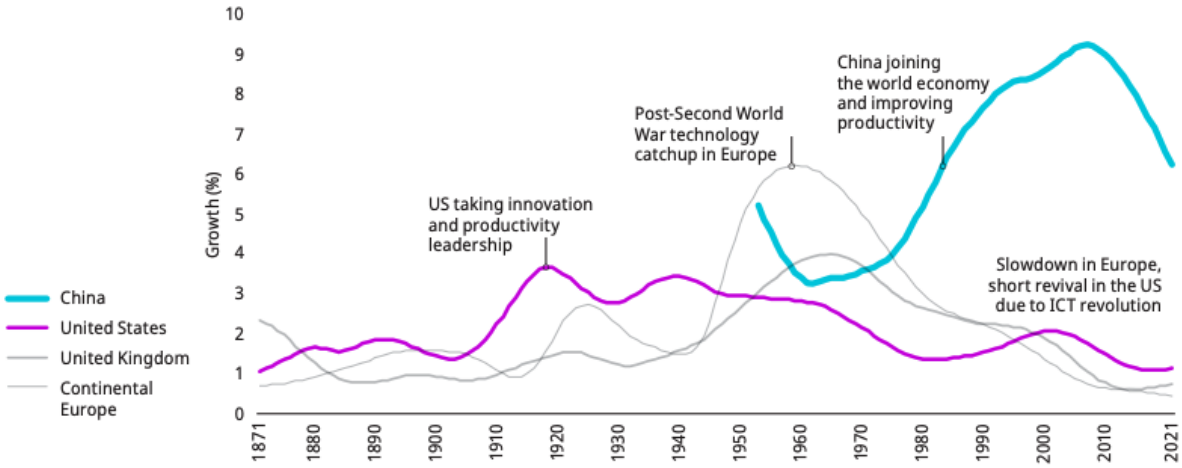
Notes: "Core" corresponds to Western Europe and its offshoots (Australia, Canada, New Zealand and the United States) with Japan. "Periphery" corresponds to the world, excluding the "core" countries.

- Unemployment Rates: World Bank https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS
- poteinail data variables, from **Global Innovation Index 2022: What is the future of innovation–driven growth? page 68**

Is the persistent productivity slowdown getting worse?

After the 1970s, a period of sustained slowdown in productivity growth began (Figure 14; see also GII 2022 Expert Contributions from van Ark and Fleming; Petropolous). Before then, productivity growth had been stimulated by the aforementioned innovation waves: the United States of America took the innovation and productivity lead in the 20th century, with the post-Second World War period especially fruitful, as technology diffused out from the more advanced United States to reach Europe and later Japan and the Republic of Korea.

Figure 14 Labor productivity growth, 1871–2021



About how many observations? How many predictors?

- depends on the answer of the question i have put at the end

What types of variables will you be working with?

- continuous variables

Is there any missing data? About how much? Do you have an idea for how to handle it?

- yes and would like to discuss about this as well. Missing years and etc..

  Thinking to either just not count those missing years, or still count them but not put in consideration. This depends on the dataset im collecting.

An overview of your research question(s)

What variable(s) are you interested in predicting? What question(s) are you interested in answering?

- How does technology advancement impact unemployment rates
- **Potential sub questions**: Does the relationship between technology advancement and unemployment differ across countries or regions, and if so, why?

Name your response/outcome variable(s) and briefly describe it/them.

- Responce: Unemployment rate (The unemployment rate represents the percentage of the total labor force that is unemployed but actively seeking employment and willing to work.)

Will these questions be best answered with a classification or regression approach?

- Regression

Which predictors do you think will be especially useful?

- R&D Expenditure

Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.

- Predictive

Your proposed project timeline

When do you plan on having your data set loaded, beginning your exploratory data analysis, etc?

Provide a general timeline for the rest of the quarter.

-

Week of April 17:

- Start exploratory data analysis (EDA) over the weekend.
- Collect, clean, and preprocess data.
- Identify missing data and decide on appropriate imputation strategies.
- Visualize data and identify trends and relationships between variables.
  such as plotting

Week of April 24:

- Continue EDA and finalize data preprocessing.
- Review literature and research methods related to topic.
- Develop hypotheses based on EDA findings.

Week of May 1:

- Select appropriate machine learning algorithms or time series models for my project.
- Split data into training and testing sets.
- Train and evaluate models using cross-validation.

Week of May 8:

- Fine-tune models and select the best-performing one.
- Conduct inferential analysis to estimate causal effects and test hypotheses.
- Perform feature importance analysis to understand the role of different predictors.

Week of May 15:

- Interpret model results and draw conclusions based on my findings.
- Identify potential limitations and biases in analysis.
- Brainstorm policy recommendations and potential future research directions.

Week of May 22:

- Start writing research paper or project report.
- Organize my findings and results in a coherent manner.
- Create clear and informative visualizations to support my analysis.

Week of May 29:

- Continue writing paper or report.
- Revise and edit paper or report based on feedback from instructors.
- Ensure that paper or report follows the appropriate structure and format.

Week of June 5:

- Finalize paper or project report.
- Ask some final questions to professor

Tuesday, June 13, 2023:

- Submit final project and any accompanying materials.

Any questions or concerns

Are there any problems or difficult aspects of the project you anticipate?

Any specific questions you have for me/the instructional team?

- I am unsure whether I should focus my data collection and prediction efforts solely on the USA or expand the scope to include the whole world. My concerns arise from the potential influence of cultural, traditional, and societal differences across countries. Could you provide guidance on selecting an appropriate scope for this project and addressing the challenges posed by cross-country variations?

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.