**APY313 Data Practical**

**Learning Objectives:**

* Get familiar and comfortable creating R Markdown reports.
* Be able to implement R code and quantitative methods (data transformation, visualization, and analysis) with real data of your choosing.
* Learn the basics of creating reproducible data science.
* Have familiarity with software and tools for data science.

**Instructions:**

* Create a “scientific” report using R Markdown with an accompanying output file (PDF, HTML, GitHub markdown). The output file in “knitted” from the working file, i.e., the Rmd file.
* The submission must be both the R Markdown file and the output file, plus all the data files, figures, etc.
* The report itself will be on some topic(s) that I am interested in (pretty much anything I want). Some examples Steve provided deal with using quantitative analysis and data visualization on a master’s thesis, scientific paper, data journalism of some sort, etc.

<https://doi.org/10.1371/journal.pone.0245522>

<https://github.com/urban-m/elev/tree/master/phylogenetic_study>

<https://github.com/urban-m/elev>

<https://en.wikipedia.org/wiki/Data_journalism>

<https://pudding.cool/>

<https://pudding.cool/projects/vocabulary/index.html>

**Requirements (Grading):**

* The report must be in R Markdown (can use Python using Jupyter notebook or Google Colab).
* The report, e.g., README.Rmd must include a header with metadata (e.g., title, author, date, output format) that successfully knits into an output format (PDF, HTML, GitHub markdown).

<https://en.wikipedia.org/wiki/Metadata>

* The report must use R Markdown syntax for formatting, e.g., headers and sub-headers, text formatting (italics, bold), bullet lists, pictures, etc.
* The report must have a bibliography (citing sources) of references cited, and it must be compliant with a specific citation style (I’ll use MLA).
* The report must contain R code blocks, i.e., you will have to implement code in R including loading and displaying the data and transforming (“wrangling”) data.

<https://github.com/bambooforest/APY313/blob/main/3_data>

<https://github.com/bambooforest/APY313/blob/main/4_data_wrangling>

* The report must use some dataset(s). I have to describe the data, e.g., in terms of its R vs statistical data types. Knowing the data types will help visualize the data.
* The report must implement some data visualization(s) of the dataset(s).

<https://github.com/bambooforest/APY313/blob/main/5_data_visualization>

<https://github.com/bambooforest/APY313/blob/main/6_data_visualization>

* I must ask a question, propose a hypothesis, and model my data.

<https://github.com/bambooforest/APY313/blob/main/7_data_modeling>

* The data model(s) and quantitative method(s) must fit the questions(s) that I’m asking.

<https://github.com/bambooforest/APY313/blob/main/8_Linear_Models_I>

<https://github.com/bambooforest/APY313/blob/main/9_Linear_Models_II>

<https://github.com/bambooforest/APY313/blob/main/10_Dimensionality_reduction_clustering_I>

<https://github.com/bambooforest/APY313/blob/main/11_Dimensionality_reduction_clustering_II>

<https://github.com/bambooforest/APY313/blob/main/9_time_series>

<https://github.com/bambooforest/APY313/blob/main/12_machine_learning>

* There must be an introduction that introduces my work.
* To conclude, I have to be able to present and discuss my results in a discussion or conclusion section.
* For overall structure, use the IMRAD model.

<https://github.com/bambooforest/IntroDataScience/tree/main/2_writing_scientific_reports#scientific-reports-why>

**Data:**

* To find the right data, I will first try googling using key words like “open data” or “raw data”.

<https://zenodo.org/>

<https://osf.io/>

<https://eosc-portal.eu/>

<https://www.nature.com/sdata>

<https://lrec2020.lrec-conf.org/en/shared-lrs/>

<https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html>

**Data Practical Examples:**

<https://github.com/gyhhuang/cbb_freshman_project/blob/main/CBB%20Project%20Text.pdf>

<https://github.com/ymijsters/My_Data_Practical>

<https://github.com/alex-ro96/ADHD_in_women>

<https://github.com/DidierYourassoff/IntroDataScience>

**Consensus AI Results:**

Given the flexibility and requirements of your project, it's a great opportunity to explore a topic that not only interests you but also showcases your skills in data science and computer science. Below are some topic ideas along with sources where you can find datasets in CSV format. Each of these topics is broad enough to allow you to ask meaningful questions and apply data modeling techniques, which could make your project stand out:

**Climate Change Analysis:**

* **Topic Idea**: Analyzing global temperature changes over time and their impact on sea levels or Arctic ice melt.
* **Data Source**: The NASA Climate Change and Global Warming website offers various datasets, including global temperature, sea ice extent, and carbon dioxide levels. <https://climate.nasa.gov/>

**Stock Market Prediction:**

* **Topic Idea**: Predicting stock market trends using historical stock prices and applying machine learning models.
* **Data Source**: Yahoo Finance provides historical stock price data that can be downloaded in CSV format. Use APIs or manual download for specific stocks of interest. <https://finance.yahoo.com/>

**COVID-19 Impact Analysis:**

* **Topic Idea**: Examining the impact of COVID-19 on various sectors such as healthcare, employment, and education.
* **Data Source**: Our World in Data provides an extensive COVID-19 dataset, including cases, deaths, vaccination rates, and more, which is regularly updated. <https://ourworldindata.org/coronavirus>

**Social Media Sentiment Analysis:**

* **Topic Idea**: Analyzing the sentiment of tweets or Reddit posts regarding a particular event or trend to gauge public opinion.
* **Data Source**: While direct CSV datasets might be harder to come by, Twitter and Reddit APIs allow for the collection of data which can then be exported to CSV. Tools like Twint can be used for Twitter data scraping without needing API access.

**E-commerce Sales Analysis**:

* **Topic Idea**: Analyzing trends in e-commerce sales, customer behavior, and product reviews.
* **Data Source**: The UCI Machine Learning Repository has multiple datasets related to e-commerce, including the Online Retail dataset, which contains transactions from a UK-based online store. <https://archive.ics.uci.edu/ml/index.php>

For your project, consider these steps: Choose a topic that not only interests you but also has relevance in today's world, enhancing your project's appeal for internships and job applications. Once you've selected your topic, formulate a specific question, or set of questions you want to answer. This will guide your data analysis and model building. Download the dataset in CSV format from the suggested sources or use APIs to collect data if necessary. Use R Markdown to create your scientific report, incorporating R code blocks for data loading, wrangling, and visualization. Ensure you follow the IMRAD model and MLA citation style as required. Remember, the key to a successful data science project is not just in conducting complex analyses but also in asking meaningful questions and being able to clearly communicate your findings and their implications. This approach will demonstrate your analytical and communication skills, making your project a strong addition to your resume.

Introduction:

The relationship between the escalating concentrations of greenhouse gases, specifically carbon dioxide (CO₂) and methane (CH₄), and the increases in global average temperatures support contemporary understanding of climate dynamics. This connection not only highlights the anthropogenic impacts on climate change but also serves as a critical point for scientific questions aimed at explaining the effects and potential mitigation strategies for global warming.

Historical and modern scientific efforts have greatly contributed to the understanding of this relationship. Eunice Foote and, later in the 19th century, John Tyndall, with their pioneering experiments, both proved the heat-trapping characteristics of CO₂, as part of the basic knowledge toward the greenhouse effect, which is necessary for modern-day climatic science (Armstrong et al., 2018). In the 20th century, scientists, including Mann et al. (1998), further attempted to provide an explanation for the dominant role of the gases in recent climate variability and to confirm the continual impact that had already been observed in earlier research by using paleoclimatic information to link changes in greenhouse gas concentrations with large-scale temperature patterns.

The most recent studies, for instance, Hansen et al. (2006) and Mann et al. (1998), have not only upheld the early findings with great analysis but have continued to argue out how these gases influence modern climatic patterns. Some of these include changes in SSTs and their link to key climate events such as the El Niño. His work, through the documentary, indeed underscores near effects of these gases and points out to the constant increase in the temperatures around the world, which is in line with past predictions and models that sought to establish the trajectory of climate change.

They increasingly report that the increase of world temperatures is very much related to rising sea levels—one of those things they are seeing more often in climate studies of recent years. Mitchell (1989) pointed out that the enhanced greenhouse gas concentrations have warmed not only the atmosphere but have also contributed to its thermal expansion, melting ice caps, which have emerged as major factors for the increased observed rise in sea levels over the past century.

Building on these foundational experiments and studies, the present research aims to delve deeper into the correlations and causal relationships between the increases in specific greenhouse gases with global temperature trends. Using context from these studies and modern technological and methodological advances, this piece of work contributes to having a more cohesive understanding of the mechanisms involved in climate change by potentially confirming or challenging past findings.