

This problem set consists of 3 pages. Try to answer all questions. For each question, each sub-question carries the same weight.

All questions should have an answer in plain text – in some cases possibly very short. For many questions it is also relevant to show the R code you have used. Be sure to make the distinction between plain text and code clear, e.g. by using different fonts.

In the grading, most emphasis is put on a correct and clear answer. Some weight is also put on clean and efficient coding.

Data for questions I. and II. can be accessed at

<https://www.dropbox.com/sh/xuqjd8jac7zvqt7/AACCDqV0InL1emya5fGdIMwUa?dl=0>

I. Democracy in the world (weight 40%)

In this question we are going to study measures of democracy in the world. To study democracy, political scientists (and others) have been constructing indices of democracy. Here a country is assigned a score at each point in time where higher values mean a more democratic regime.

Initially we will be using data from the V-Dem project.¹ The attached file `vdem.csv` is a comma separated file with data for a number of countries and years. Our measure of democracy is the variable `v2x_libdem`, a measure of liberal democracy.

- Read the data into a tibble `vdem`. What is the range of years for the countries in the data set? How many countries were present in 2021?
- Extract the data for Norway and the US (country codes NOR and USA). Draw a single time series plot with the value of the measure of democracy for each of the two countries.
- Use the data from 1950 onwards for all countries available, and compute the median democracy score for each year. Next, find the democracy scores at the 25% and 75% percentiles for each year. Show all three values in a time series diagram.
- Transform the tibble `vdem` into a tibble where each row is a country and the remaining columns contain the democracy score for the years 1980, 1990, ..., 2020. These columns should be named `demo_1980` etc. Hence a structure like this one:

country_name	demo_1980	demo_1990	demo_2000	demo_2010	demo_2020
Mexico	0.132	0.195	0.435	0.462	0.412
Suriname	0.19	0.559	0.686	0.658	0.627
Sweden	0.879	0.884	0.881	0.889	0.879
Switzerland	0.813	0.828	0.834	0.848	0.851
⋮	⋮	⋮	⋮	⋮	⋮

¹ See <https://www.v-dem.net/> for details.

Another index of democracy is provided by Freedom House.² Their ranking of democracy for a number of countries and territories in 2021 is included as the Excel file [FreedomHouse.xlsx](#). Territories are labelled with an asterisk (*) after the name of the territory.

- e) Read the Freedom House data into an R tibble `fh`. The variable “Total Score and Status” contains both the numerical democracy score and a plain text categorization. Extract the numerical score into a numerical variable `score` and the category into a string `status`. Find the range and the mean of the democracy score.
- f) Merge all countries (but not territories) from the V-Dem and the Freedom House data. Create a scatter plot with the two democracy indices against each other.
Note: The spelling of country names varies for some countries in the V-Dem and Freedom House data

II. Ranking chocolate (weight 35%)

In this question we are going to use data from the Flavors of Cacao web site. They provide results from tasting a large number of chocolate bars. The data are provided in the accompanying tabulator delimited data file [chocolate.tsv](#).

For each chocolate bar, we have the following variables

REF	Id number
Manufacturer	Name of manufacturer
Company_Location	Country of manufacturer
Review_Date	Date bar was reviewed
Bean_Origin	Country of origin of cocoa bean
Bean_Name	Name of cocoa bean
Cocoa_Percent	Percent cocoa in the bar
Ingredients	Ingredients: B = Beans, S = Sugar, S* = Sweetener other than white cane or beet sugar, C = Cocoa Butter, V = Vanilla, L = Lecithin, Sa = Salt
Characteristics	Memorable characteristics of bar
Rating	Rating

The chocolate bars are ranked on the following scale

4.0 - 5.0	Outstanding
3.5 - 3.9	Highly Recommended
3.0 - 3.49	Recommended
2.0 - 2.9	Disappointing
1.0 - 1.9	Unpleasant

- a) Read the data into R. Make a table with the frequency of values for the ranking and show these in a suitable graph.
- b) Extract the variables `Rating`, `Cocoa_Percent` and `Review_Date` in a new tibble. Split this sample into an 80 % training and 20 % testing sample. Train a linear (OLS) model where you

² See <https://freedomhouse.org/> for details.

predict the rating with the two other variables. Explain briefly how the linear model uses the explanatory variables to predict the outcome variable.

- c) Explain what a random forest algorithm does and how it differs from the linear model you used above. Use the same data and the same split as above to train a random forest model to predict the rating with the two other variables.
- d) Use the linear and random forest models from the preceding questions to predict the rating in the testing sample. Find the prediction error and compute the mean squared error for each of the two models. Compare the performance of the two models.
- e) Explain why splitting the initial sample into a training and a testing sample is important to evaluate the model in a good way. An alternative approach would be to use the whole sample for both training and testing. Explain the weaknesses with this approach.
- f) We want to add the country of origin of the cocoa bean as an explanatory variable. To avoid challenges with small groups, we want to group all origins where there are 20 occurrences or less into an Other category. Make this initial data transformation. Train lean and random forest models with cocoa percent, review date and the set of countries of origin as the explanatory variables. Compare the performance of the two models and compare with the performance of the simpler models trained above.

III. Intertemporal consumption decisions (weight 25%)

A consumer lives for two periods. Today she has an exogenous income $Y = 10$. She must choose how much to consume today c_1 and next period c_2 . She faces an interest rate of $r = 0.05$, yielding second period consumption

$$c_2 = (1 + r)(Y - c_1).$$

Her utility function is given by

$$U(c_1, c_2) = \frac{c_1^{1-\gamma}}{1-\gamma} + \beta \frac{c_2^{1-\gamma}}{1-\gamma}$$

with $\gamma = 0.4$ and $\beta = .96$.

- a) Write a function to compute utility as a function of c_1 and c_2 . Taking the budget constraint into account, make a graph with utility as a function of c_1 .
- b) Use numerical maximization to find the optimal value of c_1 . Use your preferred maximization algorithm and explain briefly the intuition behind your maximization algorithm. Comment on how the maximum matches the graph from question a).
- c) Assume now that the return to savings is random. Specifically, 1 unit saved in the first period yields a total return R , so second period consumption is $c_2 = R(Y - c_1)$. Assume that R follow an exponential distribution with rate $\lambda = 0.9$. Use Monte Carlo integration to find the expected utility from the choice $c_1 = 5$.
- d) Use numerical maximization to find the value of c_1 that maximizes expected utility when returns to savings are random as described above.