Problem Set 9 (Extra Credit)

Keep your rendered `.pdf` to no more than 5 pages. Only provide code in the rendered `.pdf` when it is specifically requested.

Again, find the dataset summarizing some Massachusetts employment statistics called `CESReport.csv`. It was downloaded from here: https://lmi.dua.eol.mass.gov/LMI/CurrentEmploymentStatistics. Use the 136 observations and 15 variables that you obtain after carefully cleaning the data as requested in previous problem sets, removing uninformative rows and rows corresponding to 2025, and converting variables to the appropriate type.

(a) Create a dataset with $1,630$ rows and 5 columns, `Year`, `CES.Series.Code`, `Description`, `Month`, and `Jobs`, where each row corresponds to a year, month, and sector and `Jobs` contains the corresponding number of jobs. Print the code you use to do this to the rendered `.pdf`.

(b) Identify the sector with the highest average within-year variability. Print the code you use to do this to the rendered `.pdf`.

(c) Create a new variable `nMonth` that takes on numeric values for each month, i.e. `1` for January, `2` for February, etc. Print the code you use to do this to the rendered `.pdf`.

(d) For the sector with description "Retail trade," make a scatter plot plot of jobs on the y axis against your numeric variable month on the x axis, where the color of points is given by the corresponding year minus 2020. Make sure that your plot is clearly annotated and self contained, e.g. make sure that which sector you chose is clearly indicated from the plot.

(e) Using `reshape_wider`, create a data frame with 48 rows, one for each year and month, and 36 columns, one for year, one for month, and one for each value of `Description`. The entries of the columns corresponding to each value of `Description` should be the corresponding number of jobs. Print the code you use to do this to the rendered `.pdf`.

(f) Using the output from the previous part and the `cor` function, create a correlation matrix with 34 rows and columns. Note: You'll want to pay attention to arguments of the `cor`

function that determine how missing values are treated. Print the code you use to do this to the rendered `.pdf`.

(g) Find the pairs of sectors with the strongest positive correlation, the strongest negative correlation, and the weakest correlation. Print the code you use to do this to the rendered `.pdf`.

(h) Download the additional data on earnings across years and sectors available here: https://download.bls.gov/pub/time.series/ce/ce.data.02b.AllRealEarningsAE. Detailed information about this data is provided here: https://download.bls.gov/pub/time.series/ce/ce.txt. You don't need to get into the weeds understanding it, but you may find it helpful to look at as you go. Read this data into `R` and only keep rows that contain CES" in the `series_id` variable. Print the code you use to do this to the rendered `.pdf`.

(i) The `series_id` variable describes the sector in a more detailed way than our `CES.Series.Code` variable. Later digits describe finer levels of detail, and the first few digits describe the coarse sector definition that matches the level of detail we have in our original dataset. Using functions for manipulating strings, create a new variable in the data from the previous part called `CES.Series.Code` that matches the first 5 characters of `CES.Series.Code` in our original data, excluding the dash. Print the code you use to do this to the rendered `.pdf`.

(j) Aggregate the new data so that you have one observation per new coarser `CES.Series.Code` and year and the average of `value` across all observations with that `CES.Series.Code` and year. Print the code you use to do this to the rendered `.pdf`.

(k) Aggregate the original data so that you have one observation per `CES.Series.Code` and year and the sum of all `Jobs` values across all observations with that `CES.Series.Code` and `Year`. Print the code you use to do this to the rendered `.pdf`.

(l) Merge the two datasets by sector and year, keeping only sectors that appear in both. You may have to modify the `CES.Series.Code` variable in the original dataset. Print the code you use to do this to the rendered `.pdf`.

(m) Aggregate the new merged dataset so that you have one observation per `CES.Series.Code` and the average earnings and jobs variable for each `CES.Series.Code`. Print the code you use to do this to the rendered `.pdf`.

(n) Using the merged data, plot total jobs against average earnings using an x-axis from 0 to $10,000,000$. Using the `text` command in `R`, add text on top of each point that describes the sector, adjusting `cex` to make things fit.