Final Project Description

Background and Introduction

The final project for the course will require you to complete some exploratory analyses based on data from Kaggle on CPU and GPU processors over time. The link to the Kaggle page is below.

https://www.kaggle.com/datasets/michaelbryantds/cpu-and-gpu-product-data (https://www.kaggle.com/datasets/michaelbryantds/cpu-and-gpu-product-data)

I have attached the dataset to the project description on Crowdmark so that you do not have to register for an account. The data on Kaggle is comprised a single CSV file containing data on approximately 5,000 GPUs and CPUs.

```
library(tidyverse)
cpu_gpu_data<-read_csv("chip_dataset.csv")
names(cpu_gpu_data)
```

```
 [1] "ID"                   "Product"              "Type"
 [4] "Release Date"         "Process Size (nm)"    "TDP (W)"
 [7] "Die Size (mm^2)"      "Transistors (million)" "Freq (MHz)"
[10] "Foundry"              "Vendor"               "FP16 GFLOPS"
[13] "FP32 GFLOPS"          "FP64 GFLOPS"
```

The data contains basic information about the processors: type (CPU vs GPU), release date, foundry (which company made the semiconductors) and the vendor (which company built the processor). It also contains physical and performance characteristics (process size, thermal design power (TDP), die size, transistors, and frequency). We will not use the columns for FLOPS (a measure computing speed).

Objectives and evaluation

The project requires you to complete two tasks, detailed below. You should prepare a single report containing your answers to both tasks. Include the code for each task in your report, for reproducibility purposes. You may include the code as code chunks where the analyses are taking place or, if you prefer, you may include it at the end (although the code should be clearly commented so that it is clear which task each block of code corresponds to).

The completion of each task is worth 35 points. The quality of presentation will also be worth 30 points, i.e. clarity of explanation, plots, tables, and code.

The length of the projects will vary, depending on the number and formatting of figures and tables and the conciseness of the writing. Rather than focusing on the number of pages, I encourage students to focus on completing each task (and subtask) below to the best of their ability in the clearest and most efficient manner.

Tasks to complete

Task 1: Overall summaries of the data

The first task is to provide some exploratory data analyses and describe the distributions of the basic information and of the five physical and performance characteristics and their associations.

a. For CPU's and GPU's **separately**, summarize **numerically** and **graphically** each of the five physical and performance characteristics using appropriate summary measures and plots.
   Compare the distributions of GPU's and CPU's for each measure, i.e. how are they most similar and how are they most different in terms of central location, spread, skew and outliers.
   Be sure to report on any missing observations as well.

b. Do you see any strong association between the number of processors released by the vendors and the foundries, i.e. do some vendors release semiconductors exclusively from one foundry or vice versa? Does this association depend on whether they are CPU's or GPU's? Use both numerical and graphical summaries to support your conclusions. **Note:** You can (should?) collapse some of the foundries with smaller counts into a single level to facilitate interpretation.

c. Does the association between Die Size and Thermal Design Power depend on Type? Explain your answer both numerically and graphically.

Task 2: Change over time

a. Compare the foundries (grouped in the same way in Task 1) in terms of the total number of processors relased by year graphically and numerically. Do the same comparison for the vendors. Describe the trends that you see over time. **Hint:** make sure you handle the Release Date column properly.

b. Moore's Law: Moore originally posited that the number of transistors per microchip would double every two years. Assess whether or not this is true by computing what you would expect to see if Moore's law held for CPU's and GPU's separately and then comparing to what you observe in the data both numerically and graphically (*Hint:* you can use the average number of transistors per processor per year).