VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE & ENGINEERING

PROBABILITY AND STATISTICS

THE EVOLUTION OF COMPUTER PROCESSORS: A STATISTIC OF COMMON PROPERTIES

Supervisor: Nguyen Thi Mong Ngoc, PhD
Students: Chau Dang Minh - 2013748

Ho Chi Minh City, April 2024

EVALUATION

N.O.	Student	ID	Works	Completed
1	Chau Dang Minh	2212287	Dataset overview Preprocessing	100%
2	Ha Khoi Nguyen	2212287	Descriptive statistics	100%
3	Nguyen Thi Mai Anh	2210103	Theories Slides and Presentation	
4			Inferential statistics	
5			Inferential statistics	

Table of Contents

Chapter I

Introduction

Phenomena that are meaningful to humans appear not to be stochastic. In the same sense, datasets produced by humans, or nature in time circulations have insights to be analyzed, which is accounted by Statistics. Thanks to Dr. Nguyen Thi Mong Ngoc's supervision in Probability and Statistics course, we have a chance to study basic statistics within an assignment with a tiny dataset. We organized our report in the following structure

1. Overview of the dataset. In this chapter, we carefully describe in details as much as possible the dataset, specifically the properties of each instance. We also notice which features to be used for later statistical tasks.
2. Preprocessing. We process data cleaning and some computations.
3. Descriptive statistics. We calculate some qualitative features of the dataset.
4. Inferential statistics. Our problems are explicitly stated and solved.

Chapter II

Overview of the Dataset

As Computer Science students, we are assigned to analyze a dataset about computer processors, namely CPUs and GPUs. Our dataset is credited to Intel, Game-Debate, and the companies involved in producing the part. Information of CPUs and GPUs are collected separately into two files, namely `Intel_CPUs.csv` and `All_GPUs.csv`.

Chapter III

Preprocessing

1 Data Cleaning

With `RStudio`, the working directory is automatically determined. Otherwise, it can be indicated by `here` library.

Listing 3.1: Required libraries and working directory setup

```
# Libraries and options
library(dplyr)
library(here)
library(knitr)
library(kableExtra)

# Self-defined functions
source("utils.R")

# Working directory
setwd(here())
```

Now our working directory have been explicated, we can use relative paths to read the data. With `RMarkdown`, we can prettify the rendering.

Listing 3.2: RStudio data object initialization

```
  # Read the CSV file into a data frame
  cpu_data <- read.csv("dataset/Intel_CPUs.csv")
  gpu_data <- read.csv("dataset/All_GPUs.csv")

  # Inspect the CPU data
  kable(head(cpu_data), format = "html") %>%
    kable_styling()
```

Invalid cells may contain `NA`, an empty string, or other values showing us that this cell's data was not collecting correctly. At the very first step, we want to selected only columns whose the percentage of valid cells exceeds our predefined value. Then we filter out all instances with invalid features. Note that careful column selection possibly remains more instances for later tasks.

Listing 3.3: Cleaning functions

```
# Check if a cell has a valid value
is_valid <- function(value) {
  return(!is.na(value)
         & !is.null(value)
         & !value == ""
         & !value == "N/A"
         & !value == "-"
```

Product_Collection	Vertical_Segment	Processor_Number	Status	Launch_Date	Lithography	Recommended_Customer_Price	nb_of_Cores	nb_of_Threads	Pr
7th Generation Intel® Core™ i7 Processors	Mobile	i7-7Y75	Launched	Q3'16	14 nm	$393.00	2	4	1.3
8th Generation Intel® Core™ i5 Processors	Mobile	i5-8250U	Launched	Q3'17	14 nm	$297.00	4	8	1.6
8th Generation Intel® Core™ i7 Processors	Mobile	i7-8550U	Launched	Q3'17	14 nm	$409.00	4	8	1.8
Intel® Core™ X-series Processors	Desktop	i7-3820	End of Life	Q1'12	32 nm	$305.00	4	8	3.6
7th Generation Intel® Core™ i5 Processors	Mobile	i5-7Y57	Launched	Q1'17	14 nm	$281.00	2	4	1.2
Intel® Celeron® Processor 3000 Series	Mobile	3205U	Launched	Q1'15	14 nm	$107.00	2	2	1.5

Figure III.1: *First instances of CPUs data*

```
          & !value == "missing"
          & !value == "unknown")
    # Add your criteria
  }

  # Select columns with enough valid cells
  filtered_data <- function(data, valid_percentage=0.8) {
  selected_columns <- character(0)

  for (col in colnames(data)) {
    valid_count <- sum(is_valid(data[[col]]))
    total_instances <- length(data[[col]])

    if ((valid_count / total_instances) >= fill) {
      selected_columns <- c(selected_columns, col)
    }
  }

  return(data[selected_columns])
}
```

Listing 3.4: Cleaned data and selected features

```
filtered_cpu_data <- filtered_data(cpu_data, 0.4)

processed_cpu_data <-
  filtered_cpu_data[
    apply(filtered_cpu_data, 1, function(row) all(sapply(row, is_valid))), ]

selected_cpu_data <- processed_cpu_data[, c("Recommended_Customer_Price",
                                            "Product_Collection",
                                            "Launch_Date",
                                            "nb_of_Cores",
                                            "nb_of_Threads",
                                            "Processor_Base_Frequency",
                                            "Bus_Speed")]

# Adjust selected columns for your later needs

kable(head(selected_cpu_data), format = "html") %>%
kable_styling()
```

	Recommended_Customer_Price	Product_Collection	Launch_Date	nb_of_Cores	nb_of_Threads	Processor_Base_Frequency	Bus_Speed
1	$393.00	7th Generation Intel® Core™ i7 Processors	Q3'16	2	4	1.30 GHz	4 GT/s OPI
2	$297.00	8th Generation Intel® Core™ i5 Processors	Q3'17	4	8	1.60 GHz	4 GT/s OPI
3	$409.00	8th Generation Intel® Core™ i7 Processors	Q3'17	4	8	1.80 GHz	4 GT/s OPI
5	$281.00	7th Generation Intel® Core™ i5 Processors	Q1'17	2	4	1.20 GHz	4 GT/s OPI
423	$281.00	7th Generation Intel® Core™ i5 Processors	Q3'16	2	4	1.20 GHz	4 GT/s OPI
424	$281.00	7th Generation Intel® Core™ i5 Processors	Q3'16	2	4	2.50 GHz	4 GT/s OPI

Figure III.2: *First instances of selected CPUs data*

2 Data Pre-computation

Some features in our data have values that need to be reformatted for easily later sorting and analyses. Therefore, we need to gain a good understand on the features.

Listing 3.5: A processing for selected features

```
cpu_columns <- colnames(cpu_data)
gpu_columns <- colnames(gpu_data)
intersect(cpu_columns, gpu_columns)
# Output: character(0)
```

Since the data files have no common features, let us take a look at them independently.