VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF APPLIED SCIENCE

PROBABILITY & STATISTICS (MT2013)

Assignment (Semester: 231)

Predicting Intel's CPU Clock Speed Using Statistical Methods

Advisor: Phan Thị Khánh Vân, FAS - HCMUT

Students: Lê Nguyễn Gia Bảo - 2210216.
Trần Đình Đăng Khoa - 2211649.
Bùi Vũ Thiên Đăng - 2252151.
Trần Tuấn Minh Khoa - 2252365.
Nguyễn Hữu Trí - 2252842.

HO CHI MINH CITY, APRIL 2024

Contents

1 Abstract

In the rapidly advancing field of computer hardware technology, understanding and predicting the clock speed (frequency) of central processing units (CPUs) is crucial for both manufacturers and consumers. This project, **"Predicting Intel CPU Clock Speed Using Statistical Methods"**, aims to develop robust predictive models for CPU clock speed based on detailed specifications of Intel CPUs. By employing statistical techniques such as **Linear Regression** and **Analysis of Variance (ANOVA)**, this study seeks to identify the key features that significantly influence CPU clock speed.

The dataset utilized in this study comprises a comprehensive collection of Intel CPU specifications, including attributes such as the number of cores, number of threads, cache size, power consumption, and various architectural details. Data preprocessing steps involve handling missing values, normalizing data, and encoding categorical variables to ensure the dataset is suitable for rigorous statistical analysis.

To identify significant predictors of CPU clock speed, ANOVA is used to assess the impact of categorical variables, providing insights into how different CPU series and generations affect clock speeds. Linear regression is then employed to model and predict CPU clock speed based on these significant features. This method directly establishes the relationship between the dependent variable (clock speed) and the independent variables (CPU specifications).

The predictive modeling component of this project primarily relies on linear regression techniques. Linear regression provides a foundational understanding of the linear relationships between the predictors and CPU clock speed. The performance of the regression models is evaluated using key metrics such as R-squared, Mean Squared Error (MSE), and visual inspection of residual plots.

Our analysis reveals that features such as the number of cores, number of threads, cache size, and power consumption are significant determinants of CPU clock speed. The linear regression model offers valuable insights into the impact of these features on clock speed, allowing for accurate predictions based on the given specifications. While ANOVA provides supplementary information on the influence of categorical variables, it is the linear regression model that forms the core of our predictive analysis.

This project contributes to the broader understanding of CPU performance dynamics, providing a methodological framework that can be applied to other hardware components or similar predictive tasks. The findings have practical implications for manufacturers in optimizing CPU design and for consumers in making informed purchasing decisions. By leveraging statistical methods and regression analysis, this study offers a data-driven approach to predicting CPU clock speed, enhancing transparency and efficiency in the marketplace.

2 Introduction

The Central Processing Unit (CPU) is often referred to as the "brain" of the computer due to its fundamental role in executing instructions and managing the operations of other components. It processes data, performs calculations, and manages tasks, making it a critical component that directly impacts a computer's performance and efficiency. As technology continues to advance rapidly, the variety and complexity of CPUs available in the market have also increased, necessitating more sophisticated methods to evaluate and predict their clock speed (frequency).

Predicting CPU clock speed accurately is crucial for several reasons. For manufacturers, understanding the factors that influence CPU clock speed can aid in optimizing CPU design, enhancing product development, and targeting the right market segments. Accurate clock speed predictions can help manufacturers maintain a balance between performance and cost-effectiveness. For consumers, knowledge of CPU performance dynamics enables informed purchasing decisions, ensuring that they obtain the best value for their money. This is particularly important given the diverse range of CPUs available, each with different specifications and performance levels.

This report focuses on the analysis of CPU specifications to predict clock speed using a variety of statistical methods. The dataset used in this study consists of detailed specifications of Intel CPUs, one of the leading CPU manufacturers in the world. Intel CPUs are widely used in various computing devices, from desktops and laptops to servers and workstations, making them an ideal subject for this study.

The primary goal of this report is to identify the key features that significantly influence CPU clock speed and to develop predictive models that can accurately estimate clock speeds based on the identified features. To achieve this, we employ several statistical techniques, including Analysis of Variance (ANOVA) and regression analysis. Each of these methods plays a vital role in understanding the relationships between different CPU specifications and their corresponding clock speeds.

Analysis of Variance (ANOVA) is utilized to determine the impact of categorical variables on CPU clock speed. By comparing means across different groups, ANOVA helps identify whether certain categorical features, such as CPU series or generation, significantly affect clock speed. This analysis provides supplementary insights that enhance our understanding of how different CPU models and architectures impact performance.

The regression analysis forms the core of our predictive modeling approach. Linear regression is applied to provide a foundational understanding of the linear relationships between the selected features and CPU clock speed. Given the complexity of CPU performance, we also consider polynomial regression to capture potential non-linear interactions among the variables. By comparing the performance metrics of these models, such as R-squared and Mean Squared Error (MSE), we can determine the most effective model for predicting CPU clock speed.

Data preprocessing is a crucial step in this analysis, ensuring the reliability and accuracy of

our models. This involves handling missing values, normalizing data, and encoding categorical variables. Proper preprocessing enhances the quality of the dataset, making it suitable for rigorous statistical analysis and modeling.

In summary, this report aims to provide a comprehensive analysis of Intel CPU specifications to predict their clock speeds. By leveraging statistical methods and regression models, we seek to develop robust predictive models that can assist manufacturers in optimizing CPU design and help consumers make informed purchasing decisions. The findings of this study will contribute to the broader understanding of CPU performance dynamics and demonstrate the value of combining various statistical techniques to create accurate and reliable clock speed predictions.

3 Background Knowledge on Statistical Methods

3.1 Hypothesis Testing

Hypothesis testing is a fundamental statistical procedure used to make inferences about population parameters based on sample data. It is essential in various fields, including scientific research, quality control, and decision-making processes. The primary objective of hypothesis testing is to evaluate the plausibility of a specific claim or hypothesis concerning a population parameter, such as the mean or variance, which is crucial for understanding CPU clock speed determinants.

In hypothesis testing, two mutually exclusive hypotheses are formulated: the null hypothesis (H_0) and the alternative hypothesis (H_a). The null hypothesis typically represents the status quo, the baseline assumption, or the claim that the researcher wishes to test against. The alternative hypothesis represents the opposite or the alternative claim that the researcher aims to support or conclude if the null hypothesis is rejected.

The process of hypothesis testing involves the following steps:

1. Formulate the null hypothesis (H_0) and the alternative hypothesis (H_a).
2. Specify the significance level (α), which is the probability of rejecting the null hypothesis when it is true (Type I error).
3. Calculate the test statistic from the sample data.
4. Determine the critical region or the critical value(s) based on the significance level and the chosen test.
5. Compare the test statistic with the critical region or critical value(s).
6. Make a decision: Reject or fail to reject the null hypothesis.

The decision to reject or fail to reject the null hypothesis is based on the comparison between the test statistic and the critical region or critical value(s). If the test statistic falls within the critical region, the null hypothesis is rejected in favor of the alternative hypothesis. If the test statistic does not fall within the critical region, the null hypothesis is not rejected.

Types of hypothesis tests relevant to this project include:

- Tests for Means: Comparing the mean clock speeds of different CPU series or generations.
- Tests for Correlation and Regression Coefficients: Assessing the relationship between CPU specifications (e.g., number of cores) and clock speeds.

Hypothesis testing is subject to two types of errors: Type I error (rejecting the null hypothesis when it is true) and Type II error (failing to reject the null hypothesis when it is false). The significance level (α) controls the probability of committing a Type I error, while the power of the test ($1 - \beta$) represents the probability of correctly rejecting the null hypothesis when it is false, where β is the probability of committing a Type II error.

3.2 Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical method used to determine if there are statistically significant differences between the means of three or more independent groups. For this project, ANOVA helps us understand the impact of categorical variables, such as CPU series and generation, on CPU clock speeds by comparing the average clock speeds across different groups.

Key Concepts:

- Hypotheses:
  - Null Hypothesis (H_0): Assumes that there are no differences in the mean CPU clock speeds among the different groups.
  - Alternative Hypothesis (H_a): Assumes that at least one group mean is different from the others.
- Between-Group Variability: Measures the variation in CPU clock speeds between different groups, reflecting the effect of the categorical variable on clock speeds.
- Within-Group Variability: Measures the variation in CPU clock speeds within each group, reflecting natural clock speed variations among CPUs of the same group.
- F-Statistic: The ratio of between-group variability to within-group variability. A higher F-statistic indicates a greater likelihood that the observed differences between group means are statistically significant.
- P-Value: The probability of observing the data assuming the null hypothesis is true. A low p-value (typically < 0.05) suggests that the differences between group means are statistically significant.

Procedure for Conducting ANOVA:

- Categorize Data: Group the dataset based on categorical variables such as CPU Series (e.g., Intel Core i3, i5, i7) and CPU Generation (e.g., 9th Gen, 10th Gen).
- Calculate Group Means: Determine the mean CPU clock speed for each group.
- Compute Sum of Squares:
  - Total Sum of Squares (SST): Measures the total variation in CPU clock speeds.
  - Between-Group Sum of Squares (SSB): Measures the variation in CPU clock speeds between different groups.
  - Within-Group Sum of Squares (SSW): Measures the variation in CPU clock speeds within each group.
- Calculate Mean Squares:
  - Mean Square Between (MSB): SSB divided by the degrees of freedom between groups.

  - Mean Square Within (MSW): SSW divided by the degrees of freedom within groups.

- Compute F-Statistic.
- Determine P-Value: Using statistical software or F-distribution tables, find the p-value corresponding to the calculated F-statistic.
- Post-Hoc Analysis (if necessary): If ANOVA indicates significant differences, conduct post-hoc tests to identify which specific groups differ from each other.

Application in This Project:

- Group Data by CPU Series: Calculate the mean clock speed for each series.
- Perform ANOVA Test for CPU Series: Formulate hypotheses, conduct ANOVA, calculate F-statistic and p-value, and interpret results.
- Group Data by CPU Generation: Calculate the mean clock speed for each generation.
- Perform ANOVA Test for CPU Generation: Formulate hypotheses, conduct ANOVA, calculate F-statistic and p-value, and interpret results.

By using ANOVA, we can systematically evaluate the impact of these categorical variables on CPU clock speeds, providing valuable insights into performance trends and refining our predictive models.

3.2.1 Formulas (for reference)

- The total sum of squares: $\mathbf{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X})^2 = \sum_{i,j} X_{ij}^2 - \frac{X^2}{N}$.
- The treatment sum of squares: $\mathbf{SSTr} = \sum_{i=1}^{k} n(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{k} \frac{X_i^2}{n} - \frac{X^2}{N}$.
- The error sum of squares: $\mathbf{SSE = SST - SSTr}$.
- Treatment degree of freedom: $df(\mathbf{SSTr}) = k - 1$. Error degree of freedom $df(\mathbf{SSE}) = N - k = nk - k$.
- The mean square for treatment: $\mathbf{MSTr} = \frac{\mathbf{SSTr}}{k-1}$.
- The mean square for error: $\mathbf{MSE} = \frac{\mathbf{SSE}}{nk-k}$.
- If H_0 is true, then the statistic $F = \frac{\mathbf{MSTr}}{\mathbf{MSE}} \sim F_{k-1,nk-k}$: Fisher random variable. If $F > F_{\alpha,k-1,nk-k}$ we reject H_0.

3.3 Linear Regression

Linear regression is a fundamental statistical technique used to model the relationship between a dependent variable (CPU clock speed) and one or more independent variables (CPU specifications). It is widely employed to analyze and make predictions based on observed data.

Objective: To find the best-fitting straight line that describes the relationship between CPU clock speed and its specifications. This line is represented by a linear equation, which takes the following form:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

Where:

- Y is the dependent variable (CPU clock speed)
- β_0 is the intercept (the value of Y when all independent variables are zero)
- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients (slopes) associated with the respective independent variables
- $x_1, x_2, ..., x_n$ are the independent variables (CPU specifications)
- ε is the error term, representing the difference between the observed values and the predicted values

The process of linear regression involves estimating the values of the coefficients (β_0, β_1, β_2, ..., β_n) using a set of observed data points. This estimation is typically performed using the method of least squares, which aims to minimize the sum of squared differences between the observed values and the predicted values obtained from the linear equation. Linear regression models can be classified into two main types:

- Simple Linear Regression: This model involves only one independent variable and is represented by the equation $Y = \beta_0 + \beta_1 x + \varepsilon$.
- Multiple Linear Regression: This model involves two or more independent variables and is represented by the equation $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$

Procedure for Linear Regression:

1. Formulate the Model: Define the relationship between CPU clock speed and its specifications.
2. Estimate Coefficients: Use the method of least squares to estimate the values of the coefficients ($\beta_1, \beta_2, ..., \beta_n$).
3. Evaluate the Model:
    - R-Squared: Measures the proportion of variance in the dependent variable explained by the independent variables.

- P-Values: Assess the significance of each coefficient.
- Residual Analysis: Check for patterns in the residuals to validate assumptions.

4. Interpret Coefficients: Understand the magnitude and direction of the impact of each independent variable on CPU clock speed. For instance, a positive coefficient for the number of cores indicates that as the number of cores increases, the clock speed tends to increase.

5. Predict: Use the model to predict CPU clock speeds based on new values of the independent variables. The predicted values can guide decisions in CPU design and marketing strategies.

Assumptions of Linear Regression:

- Linearity: The relationship between the dependent and independent variables is linear.
- Normality of Residuals: The residuals (errors) are normally distributed.
- Homoscedasticity: Constant variance of residuals.
- Independence: Observations are independent of each other.

By understanding and applying linear regression, we can develop robust models to predict CPU clock speeds based on their specifications, providing valuable insights for manufacturers and consumers.

3.3.1 Formulas (for reference)

Covariance and Correlation

- Covariance between the random variables X and Y is:

$$\text{cov}(X,Y) = \sigma_{XY} = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

- The correlation between the random variables X and Y is:

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}, \ -1 \leq \rho_{XY} \leq 1$$

Least Square Method

A statistical procedure to find the best fit for a set of data points.
Suppose that we have n pairs of observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. We have to find the linear regression model for the data as $y_i = \beta_0 + \beta_1 x_i + \varepsilon$
The sum of the squares of the deviations of the observations from the true regression line is

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1) \rightarrow min$$

The least squares estimators $\hat{\beta}_0, \hat{\beta}_1$ must satisfy

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

The least squares estimates of the intercept and the slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

where

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \left(\sum_{i=1}^{n} x_i y_i \right) - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2 \right) - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2$$

The **fitted** or **estimated regression line** is therefore:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $e_i = y_i - \hat{y}_i$: The error in the fit of the model to the i^{th} observation y_i and is called **residual**.
- SSE $= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$: The sum of squares of the residuals.
- SST $= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = S_{yy}$: The total sum of square of the response variable.
- SSR $= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \beta_1 S_{xy}$: The sum of squares for regression.
- $r^2 = 1 - \frac{SSE}{SST} = \rho_{XY}^2$: The coefficient of determination.

Estimator of Variance: $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{SST - \beta_1 S_{xy}}{n-2}$

3.4 Remark

To predict the clock speed of CPUs, we primarily rely on **Linear Regression** as it directly models the relationship between the CPU specifications (independent variables) and the clock speed (dependent variable). **ANOVA** can be used as a supplementary method to understand the influence of categorical variables on CPU clock speeds, but it is not essential for the prediction model itself.

4 Data Pre-processing

The dataset utilized in this study is derived from a CSV file containing comprehensive specifications of Intel CPUs. This dataset serves as the foundation for our analysis, providing detailed information on various CPU features that are essential for understanding clock speed trends. In this section, we describe the dataset's structure, contents, and the relevance of its attributes to our study.

4.1 Dataset Overview

The CSV file comprises numerous rows, each representing an individual Intel CPU model. Each row contains multiple attributes describing the specifications and characteristics of the CPU. The dataset encompasses a diverse range of CPU models across different series, generations, and intended applications (e.g., consumer desktops, laptops, server-grade CPUs), ensuring a comprehensive coverage of CPU types and their respective attributes.

4.2 Data Relevance and Usefulness

The following attributes from the dataset are particularly relevant for analyzing CPU clock speeds:

- **Processor_Base_Frequency**: This attribute represents the base clock speed of the CPU, measured in gigahertz (GHz). Higher base frequencies generally indicate faster processing capabilities and overall performance.
- **nb_of_Cores**: The number of cores in a CPU impacts its ability to handle multiple tasks simultaneously. CPUs with more cores can potentially support higher clock speeds under optimal conditions, affecting overall performance trends.
- **nb_of_Threads**: This attribute represents the number of threads supported by the CPU, which is often influenced by technologies like Intel's Hyper-Threading. A higher number of threads can contribute to better resource utilization and potentially higher effective clock speeds for certain workloads.
- **TDP (Thermal Design Power)**: The TDP indicates the maximum amount of heat the CPU cooling system needs to dissipate. This attribute is closely related to the CPU's power consumption and thermal characteristics, which can impact clock speed potential and performance.
- **Lithography**: This attribute refers to the manufacturing process node (e.g., 14nm, 10nm) used in producing the CPU. Advancements in manufacturing processes can lead to higher clock speeds and improved efficiency in newer CPU generations.

Our objective is to analyze historical trends in CPU clock speeds using these attributes. Understanding how clock speeds have evolved across different CPU generations, architectural improvements, and technological shifts is crucial for identifying patterns and making informed predictions. By focusing on these critical attributes, we aim to uncover insights into the factors

driving CPU clock speed improvements and contributing to the overall technological progress in CPU development.

4.3 Load Data

In this section, we load and clean the data for our analysis.

```
# Load necessary library
library (dplyr)

# Importing data
intel_cpu <- read.csv ("~/Downloads/Intel_CPUs.csv")

# Specify the desired columns
desired_cols <- c ("Processor_Base_Frequency", "nb_of_Cores", "nb_of_
Threads", "TDP", "Lithography")

# Subset the data to only include the desired columns
intel_cpu_subset <- intel_cpu %>% select (one_of(desired_cols))

# Remove rows with base frequency measured in MHz
intel_cpu_subset <- intel_cpu_subset %>%
    filter (!grepl ("MHz", Processor_Base_Frequency))

# Convert necessary columns to appropriate types if they are not
# Removing 'GHz' and converting Processor_Base_Frequency to numeric
intel_cpu_subset$Processor_Base_Frequency <-
    as.numeric (gsub (" GHz", "", intel_cpu_subset$Processor_Base_
Frequency))

# Additional conversion to ensure nb_of_Cores, nb_of_Threads, TDP, and
Lithography are numeric where applicable
intel_cpu_subset$nb_of_Cores <-
    as.numeric (intel_cpu_subset$nb_of_Cores)
intel_cpu_subset$nb_of_Threads <-
    as.numeric (intel_cpu_subset$nb_of_Threads)
intel_cpu_subset$TDP <-
    as.numeric (gsub (" W", "", intel_cpu_subset$TDP))
intel_cpu_subset$Lithography <-
    as.numeric (gsub (" nm", "", intel_cpu_subset$Lithography))
```

4.4 Handle Missing Values

In this analysis, we focus on specific attributes of Intel CPUs that directly impact their clock speeds: Processor_Base_Frequency, nb_of_Cores, nb_of_Threads, TDP, and Lithography. It is essential to ensure the integrity and accuracy of these data points for meaningful analysis and reliable model predictions. **Here's why using the median for imputation is not suitable for these attributes and why removing rows with missing values is a better approach**:

- Processor_Base_Frequency:
  - Contextual Integrity: The base frequency is a critical performance metric measured in GHz. Imputing a median value may introduce inaccuracies because the actual base frequencies are often specific to CPU models and designs. These values are precise and need to reflect the true performance characteristics of the CPU.
  - Range and Units: Base frequencies vary within specific ranges, and a median value might not accurately represent the operational characteristics of the CPUs, especially when considering variations in architecture and generation.
- nb_of_Cores and nb_of_Threads:
  - Specific Configuration: The number of cores and threads is an intrinsic property of a CPU model, determined by its design and intended use. These are exact numbers that are integral to the CPU's capability. Imputing a median would mean assigning an arbitrary number of cores or threads that do not exist in the actual hardware configurations, leading to misleading results.
  - Consistency: Each CPU model has a specific number of cores and threads. Removing rows with missing values ensures we only analyze data with complete and accurate configurations.
- TDP (Thermal Design Power):
  - Thermal Characteristics: TDP is a precise measure of the maximum heat a CPU can generate under typical load, directly influencing cooling solutions and performance. Using a median TDP might not accurately reflect the thermal design considerations specific to different CPU models.
  - Power Management: Different CPUs have different power and thermal characteristics. It's crucial to work with exact TDP values to maintain the accuracy of the analysis.
- Lithography:
  - Manufacturing Process: Lithography measures the process technology node (e.g., 14nm, 10nm). These are fixed and precise values associated with the manufacturing process of the CPU. A median value would not accurately represent any actual process node and could distort the analysis of technological trends.

```r
# Remove rows with missing values in the specific columns
cols_to_check <- c ("Processor_Base_Frequency", "nb_of_Cores", "nb_of_
Threads", "TDP", "Lithography")
intel_cpu_subset <- intel_cpu_subset[complete.cases (intel_cpu_subset[,
cols_to_check]), ]

# Display the resulting subset
print (intel_cpu_subset)
```

4.5 Handle Outliers

Handling outliers is essential for maintaining the integrity and accuracy of our data analysis. By identifying and removing outliers, we can ensure that our statistical analyses and predictive models are not unduly influenced by anomalous data points. This leads to more reliable insights and better overall model performance.

```r
# Identify Outliers Function
identify_outliers <- function (x)
{
    Q1 <- quantile (x, 0.25, na.rm = TRUE)
    Q3 <- quantile (x, 0.75, na.rm = TRUE)
    IQR <- Q3 - Q1
    lower_bound <- Q1 - 1.5 * IQR
    upper_bound <- Q3 + 1.5 * IQR
    return (x < lower_bound | x > upper_bound)
}

# Apply the function to identify outliers for each relevant column
outliers <- sapply (intel_cpu_subset[, cols_to_check], identify_outliers
)

# Combine the outliers into a single logical vector indicating any row
with an outlier
combined_outliers <- apply (outliers, 1, any)

# Print number of outliers in each column
print (colSums (outliers))

# Remove rows with outliers
intel_cpu_cleaned <- intel_cpu_subset[!combined_outliers, ]

# Verify the changes
summary (intel_cpu_cleaned)
```

4.6 Feature Scaling/Normalization

5 Descriptive Statistic

5.1 Summary Statistics

5.2 Distribution and Histograms

5.3 Boxplots for Outliers

5.4 Correlation Matrix

6 Inferential Statistics

6.1 Multiple Linear Regression to Predict CPU Clock Speed

6.2 Result

7 Conclusion

8 References

[1] Abdullah Al Sefat, G.M. Rasiqul Islam Rasiq, Nafiul Nawjis, and Sk Mehedi. *CPU Performance Prediction Using Various Regression Algorithms*, pages 163–171. 01 2021.

[2] Sheldon M Ross. *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press, 2020.

[3] Peter Bruce and Andrew Bruce. *Practical Statistics for Data Scientists*. O'Reilly, 2017.

[4] Douglas C Montgomery and George C Runger. *Applied Statistics and Probability for Engineers*. Wiley, 6th edition, 2006.

[5] William Stallings. *Computer Organization and Architecture*. Prentice Hall Professional Technical Reference, 6th edition, 2002.

[6] Nguyễn Tiến Dũng (chủ biên) và Nguyễn Đình Huy. *Xác suất - Thống kê & Phân tích số liệu*. Đại học Quốc gia TP. Hồ Chí Minh, 2019.

[7] Phan Thị Khánh Vân. Lecture notes on Probability and Statistics, LMS - HCMUT. https://lms.hcmut.edu.vn, 2024. Accessed: May 2024.

[8] Github. Hands-On Programming with R. https://rstudio-education.github.io/hopr/basics.html. Accessed: May 2024.

[9] [Học một chút]. Thống kê trong R-Studio. https://youtube.com/playlist?list=PL9XZJEFXvG9E9HJn2n1Gu_TLAsVPc1MkJ&feature=shared, 2021. Accessed: May 2024.

[10] Intel. What is Clock Speed? https://www.intel.com/content/www/us/en/gaming/resources/cpu-clock-speed.html. Accessed: May 2024.

[11] ILISSEK. Computer Parts (CPUs and GPUs). https://www.kaggle.com/datasets/iliassekkaf/computerparts, 2017. Accessed: April 2024.