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

PROBABILITY & STATISTICS (MT2013)

Assignment Semester: 231

Predicting Intel CPU Prices 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 price of central processing units (CPUs) is crucial for both manufacturers and consumers. This project, "Predicting Intel CPU Prices Using Statistical Methods," aims to develop robust predictive models for CPU prices based on detailed specifications of Intel CPUs. By employing a variety of statistical techniques including Analysis of Variance (ANOVA), Pearson correlation coefficient, and regression analysis, this study seeks to identify the key features that significantly influence CPU prices.

The dataset utilized in this study comprises a comprehensive collection of Intel CPU specifications, including attributes such as clock speed, number of cores, number of threads, cache size, and power consumption. 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 prices, ANOVA is used to assess the impact of categorical variables, while the Pearson correlation coefficient measures the strength and direction of the relationships between continuous variables and CPU prices. These statistical methods help in narrowing down the most influential features that contribute to variations in CPU prices.

The predictive modeling component of this project employs both linear and polynomial regression techniques. Linear regression provides a foundational understanding of the linear relationships between the predictors and CPU prices. However, given the complexity of CPU pricing, polynomial regression is also applied to capture more intricate, non-linear interactions among the variables. The performance of these 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 clock speed, number of cores, number of threads, and cache size are significant determinants of CPU prices. The linear regression model offers valuable initial insights, but the polynomial regression model significantly enhances prediction accuracy by accounting for non-linear relationships. The results underscore the importance of considering complex interactions among CPU specifications when predicting prices.

This project contributes to the broader understanding of CPU pricing 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 pricing strategy development 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 prices, 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 pricing.

Predicting CPU prices accurately is crucial for several reasons. For manufacturers, understanding the factors that influence CPU prices can aid in developing competitive pricing strategies, optimizing production costs, and targeting the right market segments. Accurate price predictions can help manufacturers maintain a balance between profitability and market share. For consumers, knowledge of CPU pricing 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 price points.

This report focuses on the analysis of CPU specifications to predict prices 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 prices and to develop predictive models that can accurately estimate these prices based on the identified features. To achieve this, we employ several statistical techniques, including Analysis of Variance (ANOVA), Pearson correlation coefficient, and regression analysis. Each of these methods plays a vital role in understanding the relationships between different CPU specifications and their corresponding prices.

Analysis of Variance (ANOVA) is utilized to determine the impact of categorical variables on CPU prices. By comparing means across different groups, ANOVA helps identify whether certain categorical features, such as CPU series or generation, significantly affect pricing. The Pearson correlation coefficient, on the other hand, measures the strength and direction of the linear relationship between continuous variables, such as clock speed, number of cores, number of threads, and cache size, and CPU prices. This analysis helps in identifying the most influential continuous variables that should be included in the predictive models.

The regression analysis forms the core of our predictive modeling approach. Linear regression is initially applied to provide a baseline understanding of the linear relationships between the selected features and CPU prices. However, given the complexity of CPU pricing, we also implement polynomial regression to capture 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 prices.

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 prices. By leveraging statistical methods and regression models, we seek to develop robust predictive models that can assist manufacturers in pricing strategies and help consumers make informed purchasing decisions. The findings of this study will contribute to the broader understanding of CPU pricing dynamics and demonstrate the value of combining various statistical techniques to create accurate and reliable price 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 price 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 prices of different CPU series or generations.
- Tests for Correlation and Regression Coefficients: Assessing the relationship between CPU specifications (e.g., clock speed, cores) and prices.

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 prices by comparing the average prices across different groups.

Key Concepts of ANOVA:

- Hypotheses:
  - Null Hypothesis (H_0): Assumes that there are no differences in the mean CPU prices 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 prices between different groups, reflecting the effect of the categorical variable on prices.
- Within-Group Variability: Measures the variation in CPU prices within each group, reflecting natural price 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 price for each group.
- Compute Sum of Squares:
  - Total Sum of Squares (SST): Measures the total variation in CPU prices.
  - Between-Group Sum of Squares (SSB): Measures the variation in CPU prices between different groups.
  - Within-Group Sum of Squares (SSW): Measures the variation in CPU prices 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 price 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 price 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 pricing, providing valuable insights into pricing trends and refining our predictive models.

Formulas (for reference):

- The total sum of squares: **SST** $= \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X})^2 = \sum_{i,j} X_{ij}^2 - \dfrac{X^2}{N}$.
- The treatment sum of squares: **SSTr** $= \sum_{i=1}^{k} n(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{k} \dfrac{X_i^2}{n} - \dfrac{X^2}{N}$.
- The error sum of squares: **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: $\mathbf{MSTr} = \dfrac{\mathbf{SSTr}}{k-1}$.
- The mean square for error: $\mathbf{MSE} = \dfrac{\mathbf{SSE}}{nk-k}$.
- If H_0 is true, then the statistic $F = \dfrac{\mathbf{MSTr}}{\mathbf{MSE}} \sim F_{k-1,nk-k}$: 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 price) 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 price 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
- 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
- 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 price 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 price.
5. Predict: Use the model to predict CPU prices based on new values of the independent variables.

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 these statistical methods, we can develop robust models to predict CPU prices based on their specifications, providing valuable insights for manufacturers and consumers.

3.4 Remark

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

4 Data Pre-processing

The dataset used in this study is sourced from a CSV file containing detailed specifications of Intel CPUs. This dataset serves as the foundation for our analysis, providing a comprehensive overview of various CPU features necessary for predicting prices. In this section, we describe the structure and contents of the dataset, including the types of features available and their relevance 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 that describe the specifications and characteristics of the CPU. The dataset includes a diverse range of CPU models, spanning different series, generations, and intended applications (e.g., consumer desktops, laptops, and server-grade CPUs). This diversity ensures that our analysis covers a broad spectrum of CPU types and their respective price points.

4.2 Data Relevance and Usefulness

The attributes in the dataset are crucial for predicting CPU prices because they encapsulate the key characteristics that drive market value. Below, we elaborate on the significance of each attribute in influencing CPU pricing:

- **Clock Speed**: Measured in gigahertz (GHz), clock speed is a primary performance indicator. CPUs with higher clock speeds can execute more instructions per second, thus enhancing overall performance. This makes clock speed a pivotal factor in pricing, as consumers and professionals often seek higher clock speeds for better performance.
- **Number of Cores**: The number of cores in a CPU affects its ability to handle multiple tasks simultaneously. Multi-core CPUs can manage parallel processes more efficiently, which is especially beneficial for tasks such as video editing, gaming, and running multiple applications. As a result, CPUs with more cores generally command higher prices.
- **Number of Threads**: Threads are the smallest units of processing that the operating system can schedule. CPUs with more threads can improve parallel processing capabilities, contributing to better multitasking and performance in threaded applications. This attribute is directly correlated with pricing, as higher thread counts enhance performance.
- **Cache Size**: Cache memory, typically measured in megabytes (MB), stores frequently accessed data for quick retrieval. Larger cache sizes reduce latency and improve processing speed, making CPUs with substantial cache memory more desirable and thus more expensive.
- **Power Consumption**: Measured in watts (W), power consumption impacts the efficiency and thermal performance of a CPU. CPUs that offer high performance while maintaining lower power consumption are more attractive, particularly for mobile devices

and energy-efficient systems. Efficient power consumption can justify higher prices due to lower operating costs and reduced cooling requirements.

- **TDP (Thermal Design Power)**: TDP indicates the amount of heat a CPU generates under typical load conditions. CPUs with lower TDP values can be easier to cool, which is a significant consideration for system builders looking to manage heat dissipation effectively. This can influence the price, as better thermal performance can be a premium feature.
- **Manufacturing Process**: The technology node (e.g., 14nm, 10nm) used in manufacturing a CPU affects its performance, power efficiency, and cost. Smaller manufacturing processes generally lead to more efficient and powerful CPUs, contributing to higher prices due to the advanced technology involved.
- **Release Date**: The release date provides context regarding the technological advancements and market conditions at the time of a CPU's launch. Newer CPUs often incorporate the latest technologies and improvements, justifying higher prices compared to older models.

Our objective is to develop predictive models that can accurately estimate the price of a CPU based on its specifications. Understanding the relationship between each attribute and the price is essential for creating effective models. For instance, clock speed, core count, and cache size are direct performance indicators, while power consumption and TDP relate to efficiency and usability in different environments. The manufacturing process and release date provide context about the technological and market influences on pricing.

By focusing on these critical attributes, we aim to develop models that capture the complexities of CPU pricing. This will enable us to make accurate predictions and provide valuable insights for manufacturers and consumers in the decision-making process.

4.3 Load Data

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

This line of R code is used to import a dataset from a Comma-Separated Values (CSV) file into the R environment. The `read.csv()` function is a built-in function in R that reads a CSV file and creates a data frame object from its contents. A data frame is a two-dimensional tabular data structure in `R`, where each column represents a variable, and each row represents an observation.

In this specific code: "**~/Downloads/archive/Intel_CPUs.csv**" is the file path that specifies the location and name of the CSV file to be imported, `intel_cpu` is the name assigned to the data frame object that will store the imported data.

After executing this line of code, the contents of the "**Intel_CPUs.csv**" file will be read and stored in the `intel_cpu` data frame within the `R` environment. The data frame will have the same structure as the CSV file, with columns representing variables and rows representing observations.

4.4 Explore Data

```
# The head () function is used to preview the first
# few rows of the data frame
head (intel_cpu)
```

This line calls the `head()` function and passes the `intel_cpu` data frame as an argument. By default, the `head()` function prints the first *six* rows of the given data frame or matrix.

The `head()` function is a valuable tool for data exploration and validation, especially when working with large datasets. By previewing the initial rows, you can quickly assess the structure of the data, check the column names, and ensure that the data has been imported correctly.

Inspecting the first few rows can reveal potential issues or anomalies in the data, such as missing values, incorrect data types, or unexpected values. It also provides an initial glimpse into the content and format of the data, which can inform subsequent data cleaning, transformation, or analysis steps

```
                       Product_Collection Vertical_Segment Processor_Number       Status Launch_Date Lithography
1 7th Generation Intel® Core™ i7 Processors           Mobile           i7-7Y75     Launched        Q3'16       14 nm
2 8th Generation Intel® Core™ i5 Processors           Mobile          i5-8250U     Launched        Q3'17       14 nm
3 8th Generation Intel® Core™ i7 Processors           Mobile          i7-8550U     Launched        Q3'17       14 nm
4          Intel® Core™ X-series Processors          Desktop           i7-3820 End of Life        Q1'12       32 nm
5 7th Generation Intel® Core™ i5 Processors           Mobile           i5-7Y57     Launched        Q1'17       14 nm
6     Intel® Celeron® Processor 3000 Series           Mobile             3205U     Launched        Q1'15       14 nm
  Recommended_Customer_Price nb_of_Cores nb_of_Threads Processor_Base_Frequency Max_Turbo_Frequency           Cache
1                   $393.00           2             4                 1.30 GHz            3.60 GHz  4 MB SmartCache
2                   $297.00           4             8                 1.60 GHz            3.40 GHz  6 MB SmartCache
3                   $409.00           4             8                 1.80 GHz            4.00 GHz  8 MB SmartCache
4                   $305.00           4             8                 3.60 GHz            3.80 GHz 10 MB SmartCache
5                   $281.00           2             4                 1.20 GHz            3.30 GHz  4 MB SmartCache
6                   $107.00           2             2                 1.50 GHz                                 2 MB
   Bus_Speed   TDP Embedded_Options_Available Conflict_Free Max_Memory_Size                     Memory_Types
1  4 GT/s OPI 4.5 W                         No           Yes           16 GB          LPDDR3-1866, DDR3L-1600
2  4 GT/s OPI  15 W                         No           Yes           32 GB           DDR4-2400, LPDDR3-2133
3  4 GT/s OPI  15 W                         No           Yes           32 GB           DDR4-2400, LPDDR3-2133
4 5 GT/s DMI2 130 W                         No                      64.23 GB              DDR3 1066/1333/1600
5  4 GT/s OPI 4.5 W                         No           Yes           16 GB          LPDDR3-1866, DDR3L-1600
6 5 GT/s DMI2  15 W                         No           Yes           16 GB DDR3L 1333/1600 LPDDR3 1333/1600
  Max_nb_of_Memory_Channels Max_Memory_Bandwidth ECC_Memory_Supported Processor_Graphics_ Graphics_Base_Frequency
1                         2            29.8 GB/s                   No                  NA                 300 MHz
2                         2            34.1 GB/s                   No                  NA                 300 MHz
3                         2            34.1 GB/s                   No                  NA                 300 MHz
4                         4            51.2 GB/s                   No                  NA
5                         2            29.8 GB/s                   No                  NA                 300 MHz
6                         2            25.6 GB/s                                       NA                 100 MHz
  Graphics_Max_Dynamic_Frequency Graphics_Video_Max_Memory Graphics_Output Support_4k Max_Resolution_HDMI
1                       1.05 GHz                     16 GB eDP/DP/HDMI/DVI         NA      4096x2304@24Hz
2                       1.10 GHz                     32 GB eDP/DP/HDMI/DVI         NA      4096x2304@24Hz
3                       1.15 GHz                     32 GB eDP/DP/HDMI/DVI         NA      4096x2304@24Hz
4                                                                                  NA
5                        950 MHz                     16 GB eDP/DP/HDMI/DVI         NA      4096x2304@24Hz
6                        800 MHz                               eDP/DP/HDMI         NA
  Max_Resolution_DP Max_Resolution_eDP_Integrated_Flat_Panel DirectX_Support OpenGL_Support PCI_Express_Revision
1    3840x2160@60Hz                           3840x2160@60Hz              12             NA                    3
2    4096x2304@60Hz                           4096x2304@60Hz              12             NA                    3
3    4096x2304@60Hz                           4096x2304@60Hz              12             NA                    3
4                                                                                        NA                    2
5    3840x2160@60Hz                           3840x2160@60Hz              12             NA                    3
6                                                                    11.2/12             NA                    2
  PCI_Express_Configurations_ Max_nb_of_PCI_Express_Lanes      T Intel_Hyper_Threading_Technology_
1   1x4, 2x2, 1x2+2x1 and 4x1                          10  100°C                               Yes
2   1x4, 2x2, 1x2+2x1 and 4x1                          12  100°C                               Yes
3   1x4, 2x2, 1x2+2x1 and 4x1                          12  100°C                               Yes
4                                                      40 66.8°C                               Yes
5   1x4, 2x2, 1x2+2x1 and 4x1                          10  100°C                               Yes
6                     4x1 2x4                          12  105°C                                No
  Intel_Virtualization_Technology_VTx_ Intel_64_ Instruction_Set Instruction_Set_Extensions Idle_States
1                                  Yes       Yes          64-bit       SSE4.1/4.2, AVX 2.0         Yes
2                                  Yes       Yes          64-bit       SSE4.1/4.2, AVX 2.0         Yes
3                                  Yes       Yes          64-bit       SSE4.1/4.2, AVX 2.0         Yes
4                                  Yes       Yes          64-bit          SSE4.2, AVX, AES         Yes
5                                  Yes       Yes          64-bit       SSE4.1/4.2, AVX 2.0         Yes
6                                  Yes       Yes          64-bit                SSE4.1/4.2         Yes
  Thermal_Monitoring_Technologies Secure_Key Execute_Disable_Bit
1                             Yes        Yes                 Yes
2                             Yes        Yes                 Yes
3                             Yes        Yes                 Yes
4                             Yes                            Yes
5                             Yes        Yes                 Yes
6                             Yes        Yes                 Yes
```

Console output of `head(intel_cpu)`

By executing `head(intel_cpu)`, the output will display the first *six* rows of the `intel_cpu` data frame, allowing you to visually inspect the data and make informed decisions about the next steps in the data analysis workflow.

```r
# Summary statistics
summary (intel_cpu)
```

This code snippet is used to obtain summary statistics for the dataset stored in `intel_cpu`.

The `summary ()` function in R is a versatile tool that provides a consise summary of the data, depending on the type of input it receives.

```
Product_Collection Vertical_Segment   Processor_Number      Status           Launch_Date        Lithography
Length:2283        Length:2283        Length:2283        Length:2283        Length:2283        Length:2283
Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character



Recommended_Customer_Price  nb_of_Cores      nb_of_Threads    Processor_Base_Frequency Max_Turbo_Frequency
Length:2283                Min.   : 1.000   Min.   : 1.000   Length:2283              Length:2283
Class :character           1st Qu.: 1.000   1st Qu.: 4.000   Class :character         Class :character
Mode  :character           Median : 2.000   Median : 4.000   Mode  :character         Mode  :character
                           Mean   : 4.067   Mean   : 8.728
                           3rd Qu.: 4.000   3rd Qu.: 8.000
                           Max.   :72.000   Max.   :56.000
                                            NA's   :856
   Cache             Bus_Speed             TDP            Embedded_Options_Available Conflict_Free
Length:2283        Length:2283        Length:2283        Length:2283                Length:2283
Class :character   Class :character   Class :character   Class :character           Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character           Mode  :character



Max_Memory_Size    Memory_Types       Max_nb_of_Memory_Channels Max_Memory_Bandwidth ECC_Memory_Supported
Length:2283        Length:2283        Min.   : 1.000            Length:2283          Length:2283
Class :character   Class :character   1st Qu.: 2.000            Class :character     Class :character
Mode  :character   Mode  :character   Median : 2.000            Mode  :character     Mode  :character
                                      Mean   : 2.615
                                      3rd Qu.: 3.000
                                      Max.   :16.000
                                      NA's   :869
Processor_Graphics_ Graphics_Base_Frequency Graphics_Max_Dynamic_Frequency Graphics_Video_Max_Memory Graphics_Output
Mode:logical        Length:2283             Length:2283                    Length:2283               Length:2283
NA's:2283           Class :character        Class :character               Class :character          Class :character
                    Mode  :character        Mode  :character               Mode  :character          Mode  :character



Support_4k     Max_Resolution_HDMI Max_Resolution_DP  Max_Resolution_eDP_Integrated_Flat_Panel DirectX_Support
Mode:logical   Length:2283         Length:2283        Length:2283                              Length:2283
NA's:2283      Class :character    Class :character   Class :character                         Class :character
               Mode  :character    Mode  :character   Mode  :character                         Mode  :character



OpenGL_Support PCI_Express_Revision PCI_Express_Configurations_ Max_nb_of_PCI_Express_Lanes      T
Mode:logical   Length:2283          Length:2283                 Min.   : 0.0                Length:2283
NA's:2283      Class :character     Class :character            1st Qu.:16.0                Class :character
               Mode  :character     Mode  :character            Median :16.0                Mode  :character
                                                                Mean   :20.4
                                                                3rd Qu.:32.0
                                                                Max.   :48.0
                                                                NA's   :1104
Intel_Hyper_Threading_Technology_ Intel_Virtualization_Technology_VTx_  Intel_64_          Instruction_Set
Length:2283                       Length:2283                          Length:2283        Length:2283
Class :character                  Class :character                     Class :character   Class :character
Mode  :character                  Mode  :character                     Mode  :character   Mode  :character



Instruction_Set_Extensions Idle_States        Thermal_Monitoring_Technologies  Secure_Key         Execute_Disable_Bit
Length:2283                Length:2283        Length:2283                     Length:2283        Length:2283
Class :character           Class :character   Class :character                Class :character   Class :character
Mode  :character           Mode  :character   Mode  :character                Mode  :character   Mode  :character
```

Console output of `summary(intel_cpu)`

4.5 Handle Missing Values

4.6 Handle Outliers

4.7 Feature Scaling/Normalization

5 Descriptive Statistic

6 Conclusion

References

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

[2] Nguyễn Tiến Dũng (chủ biên) and Nguyễn Đình Huy. *Xác suất - Thống kê & Phân tích số liệu*. 2019.

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

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