VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
Faculty of Computer Science and Engineering

Group 7 - CC03 — Assignment Report

Predict Memory Bandwidth using Multiple Linear Regression

Supervisors:	Dr. Phan Thi Huong	
Students:	Huynh Ngoc Van	2252898
	Pham Nguyen Hai Khanh	2252333
	Nguyen Thuy Tien	2252806
	Luu Bao Long	2252442
	Le Vu Cong Vinh	2252909

Ho Chi Minh City, January 9, 2025

Name	ID	Tasks	Adjusted Score
Huynh Ngoc Van	2252898	- **[1.1]** Context of data - **[3.1]** Data reading, **[3.2.1]** Data standardization - **[5.1]** Target, **[5.2.1]** Model definition, **[5.2.2]** Model fitting, **[5.2.3]** Result - Check content and edit report format	None
Pham Nguyen Hai Khanh	2252333	- **[1.2]** Dataset description, **[1.3]** Variable description - **[4]** Descriptive statistic - **[5.2.4.b]** Linear relationship - **[6.1]** Multiple Linear Regression	None
Nguyen Thuy Tien	2252806	- **[2.4]** Q-Q plot, **[2.5]** R-squared - **[3.2.2]** Dealing with missing data - **[5.2.4.a]** Normality	None
Luu Bao Long	2252442	- **[2]** Background - **[3.4]** Data summary - **[5.2.4.c]** Multicolinearity, **[5.3]** Testing - **[6.2]** Extension using Ridge Regression	None
Le Vu Cong Vinh	2252909	- **[2.3]** P-value - **[3.3]** Correlation coefficien between variable - **[5.2.4.d]** Homoscedasticity	None

Contents

1 Data introduction

1.1 Context of Data

The graphic processing unit, also known as GPU, is one of the most important parts in the computer. GPU is a critical component for accelerating demanding tasks, especially in graphics-intensive applications and scientific computing. Compared to CPU which handles tasks sequentially, GPU are designed for parallel processing with numerous cores that can do multiple tasks simultaneously. Additionally, GPU can achieve superior performance while consuming less energy. In essence, thanks to the powerful and efficiency of GPU, it becomes more and more significant. [6]

In this report, our group use the dataset collected by ILISSEK [10], drawing information primarily from reputable sources like Intel - a leading manufacturer of CPUs and GPUs, and Game-Debate - a well-established hardware review website. The dataset is a comprehensive collection of computer parts, encompassing intricate specifications, launch dates, and initial pricing of both CPU and GPU. Each table maintains a distinct set of entries, with features encompassing aspects such as clock speeds, peak temperatures, display resolutions, power consumption, thread count, release dates, introductory prices, die size, support for virtualization, among numerous other related parameters.

Based on the given data, the target of our report is to focus on GPU dataset and predict its memory bandwidth, which is essential for optimizing performance, resource allocation, system design, energy efficiency, and cost-effectiveness in GPU-accelerated computing systems.

1.2 Dataset Description

- **Title:** Computer Parts (GPU)
- **Source Information:** All_GPUs.csv
- **Number of Instances:** 3407
- **Number of Variables:** 34
- **Population:** GPUs

1.3 Variable Description

Variable	Data type	Description
Architecture	Categorical	Refers to the design and organization of the GPU's components
Best Resolution	Continuous	Depends on several factors, including the specific use case, the capabilities of the GPU, and the display being used
Boost Clock	Continuous	Refer to the maximum clock speed that a graphics card can reach under optimal conditions

Core Speed	Continuous	Refers to the speed at which the GPU's cores or processors operate123. These cores are responsible for rendering graphics
DVI Connection	Continuous	A type of connection used by GPUs to connect to displays
Dedicated	Categorical	A dedicated GPU, also known as a discrete graphics card, is a separate processor from the CPU and has its own dedicated memory
Direct X	Categorical	A collection of APIs developed by Microsoft that handle tasks related to multimedia, especially game programming and video, on Microsoft platforms
DisplayPort Connection	Continuous	Number of DisplayPort connections GPUs have
HDMI Connection	Continuous	Number of HDMI connections GPUs have
Integrated	Categorical	Refers to a GPU that's built into the same package as the CPU, shares everything with the CPU, including the processor package, cooling system, and system memory
L2 Cache	Continuous	Type of cache memory that is shared by all engines in the GPU, including but not limited to Streaming Multiprocessors, copy engines, video decoders, video encoders, and display controllers
Manufacturer	Categorical	Refers to the company that designs and produces the GPU
Max Power	Continuous	Refers to the maximum amount of power that the GPU can consume
Memory	Continuous	Type of high-speed memory that the GPU uses to store data it needs to perform its tasks
Memory Bandwidth	Continuous	Refers to the theoretical maximum amount of data that the GPU can transfer between its memory and other components per unit of time
Memory Bus	Continuous	Refers to the pathway that the GPU uses to access its memory
Memory Speed	Continuous	Refers to the rate at which the GPU can read and write data from its memory
Memory Type	Categorical	Refers to the type of memory technology used by the GPU to store and access data
Name	Categorical	Refers to the specific model of the GPU

Notebook GPU	Categorical	A GPU specifically designed for use in laptops or notebooks
Open GL	Continuous	A cross-language, cross-platform API for rendering 2D and 3D vector graphics
PSU	Continuous	responsible for supplying the necessary power to all the components in the system, including the GPU
Pixel Rate	Continuous	Refers to the maximum number of pixels that the GPU can render onto a screen every second
Power Connector	Categorical	Refers to the connector that allows the graphics card to draw power from the host system
Process	Continuous	Refers to the technology used to manufacture the GPU
ROPs	Continuous	Component in the graphics pipeline of a GPU
Release Date	Missing value	Refers to the date when that particular model of the GPU was officially made available to the public
Release Price	Continuous	Refers to the price at which the GPU was initially sold when it was first released
Resolution WxH	Continuous	Refers to the maximum width (W) and height (H) in pixels that the GPU can support for rendering images
SLI Crossfire	Categorical	Technologies developed by NVIDIA and AMD, respectively, that allow you to connect multiple graphics cards together to work in parallel, boosting your gaming or rendering performance
Shader	Continuous	These programs are executed for each specific stage of the graphics pipeline, transform inputs to outputs
TMUs	Continuous	Components in a GPU, responsible for manipulating bitmap images and performing texture sampling
Texture Rate	Continuous	Refers to the number of textured pixels that a graphics card can render on the screen every second
VGA Connection	Continuous	Analog interface used to connect a GPU to a display device, such as a computer monitor

Table 1: *Description of Variables*

2 Background

2.1 Multiple Linear Regression

2.1.1 Definition

Multiple linear regression is a statistical technique used to model the relationship between two or more independent variables and a single dependent variable. This method is an extension of simple linear regression, which involves just one independent variable. Multiple linear regression analyzes how various independent variables contribute to the dependent variable, and it can also accommodate interactions between different independent variables. [7]

2.1.2 Formula

The general formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Where:

- y is the dependent variable you are trying to predict.
- $x_1, x_2, \ldots, x_n$ are the independent variables.
- β_0 is the intercept of the regression line (the value of y when all x variables are 0).
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables which represent the change in the dependent variable for one unit change in the corresponding independent variable, holding all other variables constant.
- ϵ is the error term, which accounts for the variability in y that cannot be explained by the independent variables.

2.1.3 Estimation of Coefficients

The coefficients $\beta_0, \beta_1, \ldots, \beta_n$ are usually estimated using the method of least squares. This method minimizes the sum of the squared differences between the observed values and the values predicted by the model. The solution to this minimization problem typically involves matrix operations, where you calculate the vector of coefficients (β) as:

$$\beta = (X^T X)^{-1} X^T Y$$

Here, X is a matrix that includes a column of ones (for the intercept) and columns for each independent variable, and Y is the vector of observed values of the dependent variable.

2.1.4 Assumptions

Multiple linear regression analysis relies on several key assumptions:

- **Linearity**: The relationship between the independent and dependent variables is linear.
- **Independence of errors**: The residuals of the model are independent of each other.
- **Homoscedasticity**: The variance of the residuals is constant across all levels of the independent variables.

- **Normal distribution of errors**: The residuals are normally distributed (particularly important for inference regarding the coefficients).
- **No multicollinearity**: The independent variables are not too highly correlated with each other.

2.2 Ridge Regression

Ridge Regression, also known as Tikhonov regularization, is a linear regression technique used to analyze data with multicollinearity—where independent variables are highly correlated. This condition can complicate finding a stable and unique solution in standard linear regression by increasing the variance of the coefficient estimates, making them highly sensitive to model changes. Ridge Regression addresses this issue by introducing a penalty parameter λ on the size of the coefficients in the model.

2.2.1 Ridge Regression Formula

$$L(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2$$

Where:

- y_i is the actual response value.
- x_{ij} is the value of the j-th independent variable for observation i.
- β_j is the coefficient of the j-th independent variable.
- β_0 is the intercept.
- λ is the regularization parameter controlling the penalty applied to the size of the coefficients, encouraging smaller coefficients to reduce model complexity.

2.2.2 Selecting λ

Choosing the right λ value is crucial. Too high a λ can overly penalize the coefficients, potentially leading to underfitting. Conversely, too low a λ may not adequately address multicollinearity, leading to an overfitted model. λ is usually selected via cross-validation.

2.3 P-value in statistical hypothesis tests

The **P**-value is a fundamental concept in statistical hypothesis testing, representing the probability of observing a specific set of data given that the null hypothesis is true. The P-value is used in hypothesis testing to determine whether to reject the null hypothesis, with smaller P-values indicating stronger evidence against the null hypothesis. [8]

The null hypothesis (H0) is a standard component of all statistical tests. In most tests, the null hypothesis posits that there is no association between the variables of interest or no difference between groups. For example, in a two-tailed t-test, the null hypothesis asserts that there is no difference between the two groups. [8]

Consider the following example:

- Null hypothesis (H0): There is no difference in the average height between men and women.
- Alternative hypothesis (H1): There is a difference in the average height between men and women.

The decision rule with the P-value is as follows:

- If P-value (pv) $\leq \alpha$, reject H0.
- If P-value (pv) $\geq \alpha$, there is not enough basis to reject H0.

The critical value is calculated using the provided significance level (pv $\leq \alpha$) and the type of probability distribution of the idealized model. The critical value divides the area under the probability distribution curve into rejection region(s) and non-rejection region(s).

There are three types of tests: a right-tail test, a left-tail test, and a two-sided test. [8]

2.4 The Q-Q plot (Quantile-Quantile plot)

The Q-Q plot is a graphical tool used to determine if a dataset could have originated from a certain theoretical distribution like Normal or exponential. For instance, if we conduct a statistical analysis assuming our residuals follow a normal distribution, we can utilize a Normal Q-Q plot to verify this assumption. It's merely a visual inspection, not a tight proof, and thus it has a degree of subjectivity. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is breached and which data points are responsible for this breach.

In a Q-Q plot, each observation is plotted as a single dot. The x co-ordinate is the theoretical quantile that the observation should fall in if the data were normally distributed (with mean and variance estimated from the sample), and on the y co-ordinate is the actual quantile of the data within the sample. If the data are normal, the dots should form a straight line. [2]

2.5 R-Squared (R^2)

R^2, also known as the coefficient of determination, is a statistical metric that quantifies the proportion of the dependent variable's variance that is predicted by an independent variable in a regression model.

While correlation measures the degree of association between an independent and a dependent variable, R-squared measures how much of the dependent variable's variance is accounted for by the independent variable. For instance, if the R^2 of a model is 0.50, it means that around 50% of the observed variability can be accounted for by the input variables of the model.

The formula for R-Squared:

$$R^2 = 1 - \frac{\text{Unxplained Variation}}{\text{Total Variation}}$$

Computing R-squared involves a series of steps. It starts with taking the observations of the dependent and independent variables and determining the best fit line, typically using a regression model. Next, you compute the predicted values, subtract the actual values, and square the outcome. This results in a list of squared errors, which when summed up, gives the unexplained variance.

The **adjusted coefficient of determination** is the multiple coefficient of determination R2 modified to account for the number of variables and the sample size. It is calculated by:

$$\text{Adjusted}R^2 = 1 - \frac{n-1}{n-(k+1)} \times (1 - R^2)$$

3 Data preproceeding

3.1 Data reading

Firstly, we install 'tidyverse' package [14], which is designed to make it easy to install and load core packages from the tidyverse in a single command such as: 'ggplot2' for data visualisation, 'dplyr' for data manipulation, 'readr' for data importing and so on.

```
install.packages("tidyverse")
library(tidyverse)
#import .csv file
gpu_data <- read.csv("All_GPUs.csv")
View(gpu_data)
```

	Architecture	Best_Resolution	Boost_Clock	Core_Speed	DVI_Connection	Dedicated
1	Tesla G92b	NA	NA	738	2	Yes
2	R600 XT	1366 x 768	NA	NA	2	Yes
3	R600 PRO	1366 x 768	NA	NA	2	Yes
4	RV630	1024 x 768	NA	NA	2	Yes
5	RV630	1024 x 768	NA	NA	2	Yes
6	RV630	1024 x 768	NA	NA	2	Yes
7	R700 RV790 XT	1920 x 1080	NA	870	1	Yes
8	R600 GT	1024 x 768	NA	NA	2	Yes
9	Pitcairn XT GL	1920 x 1080	NA	NA	0	Yes
10	RV100	NA	NA	NA	NA	Yes
11	NV28GL A2	NA	NA	NA	2	Yes
12	Fermi GF110	1920 x 1080	NA	650	2	Yes
13	Kepler GK110	NA	NA	705	0	Yes
14	Kepler GK110	2560 x 1600	NA	706	0	Yes
15	RV200	NA	NA	NA	NA	Yes
16	GCN 1.1 Oland XT + Kaveri	1600 x 900	1100	1050	1	Yes

Figure 1: *View All_GPUs dataset*

3.2 Data cleaning

3.2.1 Data standardization

In the existing dataset, numerous columns are formatted as strings, representing values in the format "**number + unit**". To enhance data processing capabilities for future analysis, we have undertaken the task of converting these string-formatted data into numeric values.

This conversion facilitates more efficient computational operations and enables us to perform a wide range of analytical tasks with greater ease and accuracy.

Moreover, within the dataset, there exists a unique 'PSU' column structured as "**number + unit 1 + & + number + unit 2**". To facilitate more granular analysis and better represent the underlying data, we have separated this column into two distinct columns: 'PSU_Watt' and 'PSU_Amps' This segmentation allows for more precise examination of power supply unit (PSU) characteristics by isolating and quantifying the wattage and amperage components separately.

By undertaking these data transformations, we aim to optimize the dataset for comprehensive analysis while ensuring compatibility with various analytical techniques and tools.

3.2.2 Dealing with missing data

For checking the missing data in the chosen dataset, we replace N/A meaning values `None`,`""` and `\nUnknown Release Date` by `NA`.

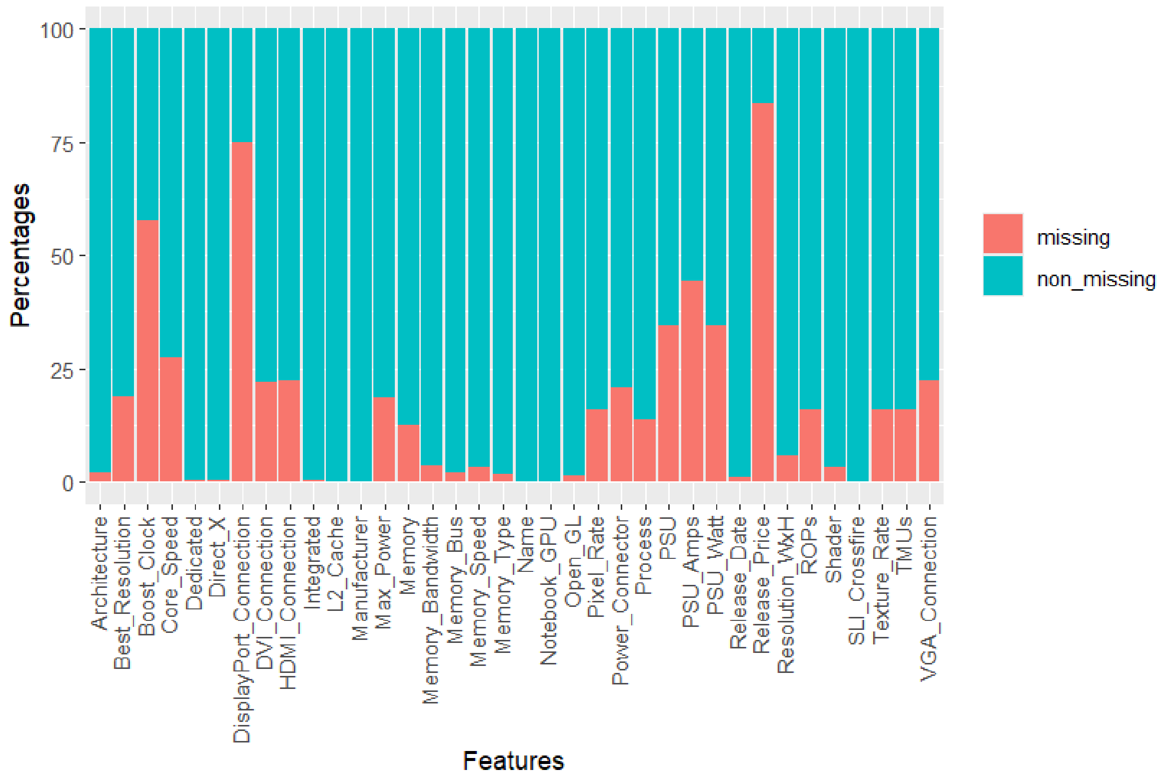


Figure 2: *Percentage of missing values in each feature*

According to Figure 2, there are some columns in which the percentage of missing data is higher than 50%, we will remove those columns. Also, base on the domain knowledge about GPU and its memory bandwidth [5] [9], we remove the columns that is not impact or relevant with memory bandwidth such as: connection type variables, release date, manufacturers,etc. Therefore, we will keep the following 15 columns features for further processing: Process, Memory bus, TMUs, PSU Amps, PSU Watt, Max Power, Open GL, Core speed, Memory speed, L2 Cache, Memory, Pixel rate, ROPs,Texture rate, and Memory bandwidth.

The sub-dataset we have chosen is mostly numeric features so there will be no high cardinality categorical features. Also, the size of this sub-dataset is medium, hence we choose the method K-Nearest Neighbors (KNN) to dealing with the missing value.

KNN is a popular method used in statistics and machine learning for dealing with missing data. It is a non-parametric, lazy learning algorithm that can be applied to both classification (need data preproceeding for categorical variables) and regression tasks. The algorithm is also highly unbiased in nature and makes no prior assumption of the underlying data. In the context of dealing with missing data, KNN imputation involves using the values of neighboring data points to estimate the missing values.

Result of imputation using this method may different depend on different value of k. Normally, for large dataset, to find out the optimal k value, we have to construct, train and test the model [1] [4]. But in the case of this medium sub-dataset with 3406 rows, we take the integer k: $k = \sqrt{3406} \approx 58$ - an relatively stable and average number.

3.3 Correlation coefficients between variables

In order to see the linear relationship between each variable, we will plot the correlation coefficient of all the possible variables using corrplot function and display these coefficients in the terminal.

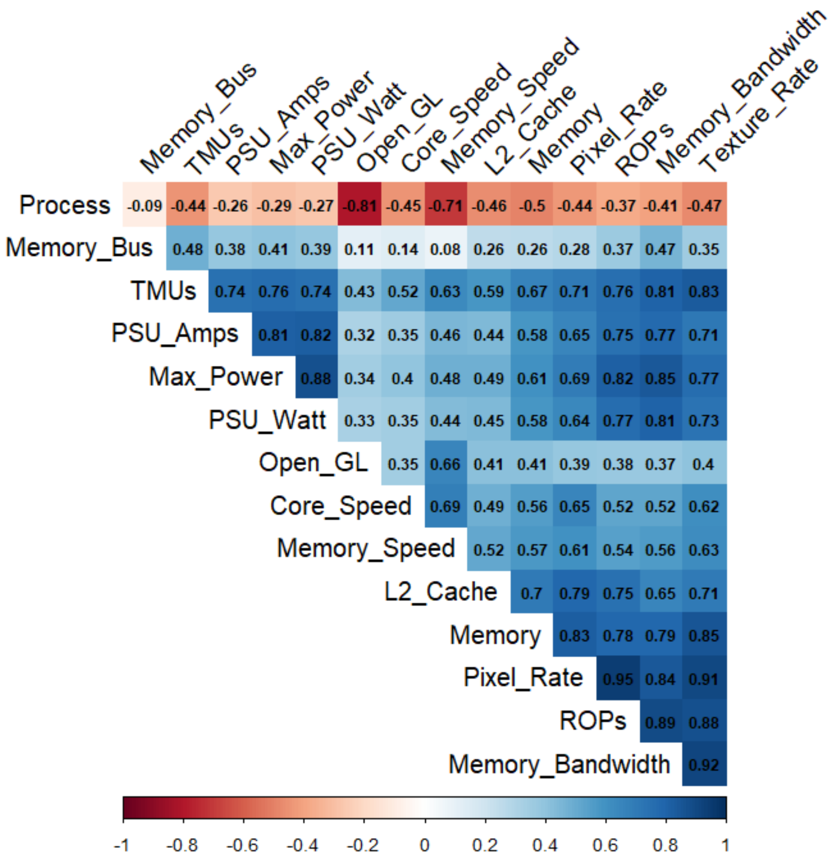


Figure 3: *Correlation coefficients plot*

However, in order to use `'corrplot'`, we must include the corrplot library. The corplot is used to draw the model that we saw a above, but we have to use cor function to calculate all the correlation coefficients.

- Close to 1 means that when one variable increases by the value of the coefficient the other tends to increase as well
- Close to -1 indicates one when increases then the other will decrease by the value of the coefficient.
- 0 means 2 variables have no relation

By analyzing the coefficients between the variables in Figure 3 , we could deduce some conclusion about factors that could have a great effect on memory which is our subject.

- First, memory seems to be greatly affected by L2 Cache, the Memory-Bandwidth, the Pixel Rate, the ROPs and the Texture Rate with high average correlation coefficients that is 0.79. Moreover,there are other factors also have noticeable contribution to the memory are Core speed, Max Power, Memory Speed, PSU AMPS, PSU WATTS and TMUs with the average value of coefficients is 0.6.
- Second, the memory bandwidth is also affected by all the other factors except for core speed, momory bus, memory speed, open GL and the process.
- In addition, the memory bus tend to be not being affected greatly by any factors.
- And finally, the memory speed is greatly affected by the core speed with the value of roughly 0.7 although the others except for process also contribute to this category.

Over the analyze above, we decided to choose variables which has the coefficients that is at least 0.7 such as the Max_Power, PSU_Amps, PSU_Watt, Pixel_Rate, ROPs, TMUs and Texture_rate and these factors will be considered carefully in our research due to their great contribution to the memory system.

3.4 Data summary

3.4.1 Max Power

The maximum power consumption data of processors ranges from as low as 1 W to as high as 780 W. The bulk of processors consume between 40 W to 100 W, which is a common power consumption range for desktop CPUs.

```
Max_Power
  1   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  37  38  39
  1   6  17   7   9   4   4  26   9  42   9  30  59   1  21  37  30  52   1   1  19  29  63   5  15  60  27  35   8   7  17   2 139   2  10  10
 40  41  42  43  44  45  46  47  48  49  50  54  55  56  57  58  59  60  61  62  64  65  66  67  68  69  70  73  74  75  77  78  80  81  83  84
 19   7   5   1  11  50   1  36  12  33  53   1  42   2   2   5  10  34   3   2  52 118  23   1   1   2  10   6   2 173   5   1  49   3   2   5
 85  86  87  88  89  90  91  94  95  96 100 101 105 106 107 108 109 110 113 115 116 120 122 124 125 127 130 134 135 138 140 141 142 143 145 146
 13  11   1   2   1  13   3   1  23   9  51   2  26  15   4  15   1  40   1  11   7 157   6   1   8   2  21   3   2   1  30   2   1   3  11   2
148 150 151 152 154 158 160 165 170 171 172 175 176 180 182 184 185 188 189 190 195 197 200 204 206 210 215 219 220 225 229 230 235 236 238 240
 13 285  10   1   1   1  24  27  61   6   1  80   1  78   3   1   5   2   1  62  21   1  67   2   2   2   7   9  45  39   2  17   2   1   2   2
244 245 247 250 251 255 256 260 265 268 270 274 275 280 289 290 294 295 296 300 302 325 340 350 360 365 375 380 385 390 400 408 410 420 430 450
 13   1   1 154   2   1   1   2  15   1   1   2  23   9   4  17   2   1   1  49   1   2   3   8   3   4  25   1   2   1   5   1   1   1   2  14
460 488 499 500 540 550 580 588 600 700 750 780
  2   1   1  26   1   8   2   1   4   1   4   1
```

Figure 4: *Summary of Max Power*

3.4.2 Memory Bandwidth

Memory bandwidths span from 1 GB/s to 1280 GB/s. There is a significant concentration of units around 1-10 GB/s, with fewer units having bandwidth greater than 100 GB/s.

```
Memory_Bandwidth
    1   1.1   1.2   1.3   1.6   1.8     2   2.1   2.3   2.4   2.7   2.8   2.9   3.2   3.6   3.7     4   4.1   4.3   4.4   4.8   5.1   5.2   5.3
    1     2     1     4     2     2     2     1     2     2     6     4     2    25     2     3     7     1     3    10     3     1     5     5
  5.6   5.8   6.3   6.4   7.2     8   8.5   8.6   8.8   9.3   9.6   9.9  10.4  10.6  10.7  11.2  12.2  12.5  12.6  12.8  13.2  13.6    14  14.3
    2     1     1    35     2    23    24     3     6     1    19     1     1     3    70    30     1     1     5    90     1     2     2     1
 14.4  14.9    15  15.2    16  16.3  17.1  17.6  19.2  20.8  21.3  22.1  22.4  23.4    24  25.3  25.6  25.9  26.6  27.2  27.9  28.5  28.8    29
   62     3     1     5    33     1    41     5     6     3    57     1    22     1     5     4   179     1     2     1     1     9   115     3
 29.6  29.9  30.1  30.4  31.7    32  32.1  34.1  35.2    36  36.8  38.4    40  40.1  41.6  42.2  43.2  44.2  44.8  46.4    48  49.6  50.2  51.2
    1    33     1     1     2    22     1    71     7     3     6    12    11    10     4     5     4     1     8     1     1     1     2    45
 53.1  54.4  57.6  57.7  59.1    60  60.2  60.8  62.7  63.4  63.6  63.9    64  64.1  65.3  65.7  67.2  69.1  70.4  71.7    72  72.1  73.6  75.5
    2     5    24     8     1     3     1     5     3     4     4     1    62     2     1     1     2     2     8     1    56     1    28     1
 76.8  79.7    80  80.2  80.3  81.6  83.2  86.4  86.6    88  89.6  89.9  91.9    96  96.1  96.2  98.4  98.5  99.2 100.8 102.4 102.7 103.2 103.7
   19     1    71    17     1     4     4    52     1     6     4     1     2    38     4     2     6     3     1     1    20     2     2     1
  104 104.5 105.6 105.7 105.8   106 106.4 107.7 108.3 108.8 108.9 111.1 111.9   112 112.1 112.2 112.3 112.9 115.2 115.5 117.6 118.4   120 120.3
   17     1     8     1    22     2     1     1     1     6     1     1     3    62    59    48     1     1    27     1     2     1     6     5
121.6 122.5 123.2 124.8 125.4   127   128 128.1 128.3 128.6 129.3 130.6 131.2 131.3 133.1 133.9 134.4 134.8   136 137.9 140.8 141.7 143.4 143.6
    4     1     2     3     1     1    26     1    18     1     2     1     2     3     1     2    36     4     1     1     9     1     1     1
  144 144.2 146.4 146.6 147.2   152 153.6 154.9   156 156.8 158.6   159   160 160.4 163.2 163.4 163.8 166.4   168 172.8 174.3   176 177.4 177.6
    8    55     1     7     4     6    51     2     4     3     3     1    36     4     1     1     1     1     1     3     1    23     2     1
179.2 182.4 185.6 188.8   192 192.2 192.3 192.4   193 194.5 194.6 195.5 196.6 196.8   197 198.4 198.7 201.4 201.6 202.2 204.8   208 211.2 211.5
   56    15     5     4    14    47    81     6     3     3     3     2     1     3     2     4     8     1     4     2     1     1    12     2
211.6 217.9   218 223.8   224 224.3 224.4 225.4 225.5 225.8 227.2 230.4 235.9   240 240.6 249.6   254   256 256.3 256.5 259.5 262.7   264 267.8
    1     1     5     2    58    10    94     1     1     1     2    13     1    30     1     3     1    65    33     1     2     1    18     1
268.8   272 272.3 273.6 281.6   288 288.4 297.6   298 298.2   304 307.2 316.8 317.2   318   320 320.3 320.8 323.3 327.9 331.8 332.8   336 336.6
    4     2     3     5     1    24    40     4     1     1     1     9     1     1     1    33    25     1     2     2     2     2    15    29
340.6 340.8 345.6 346.2 349.4   352 354.8 358.4   361 364.8   384 384.5 384.8 390.4 397.3 432.6   436 448.5 448.8 460.8 476.9   480 484.4 488.6
    4     1    12     1     1     4     2    11     1     1    25     7     2     3     1     1     1     3    21     3     1     7    38     2
489.3 493.5 494.2   502   512   528 547.2   576 576.8 595.2 633.6   640 640.5 655.9 665.6   672 673.2 681.6 691.2 692.4   720   768   769   800
    1     1     1     4     8     1     1     3     6     1     1     8     1     1     1     5     9     1     4     1     1     3     2     2
  880  1000  1024  1280
    1     1     3     1
```

Figure 5: *Summary of Memory Bandwidth*

3.4.3 Pixel Rate

Pixel rate distribution suggests a wide range of performance capabilities. The data points towards a large number of GPUs with low to moderate pixel rates, while high pixel rate capabilities are less common.

```
Pixel_Rate
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
  1 119 305 264  98 108 104  84  81  45  58  44  30  46  50  32  60 100  94  66  36  29  22  25  26  31  63  37  34  49  53  51  58  24  50  51
 36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
 33  27  57  17  41  54  23  41  19  32  17  26  15  11  18  10   6  10   3   6   1   8   9   5   6   6   3  19   5  15  22  15  12   7   3   8
 73  74  75  76  77  78  80  81  82  83  84  85  86  87  88  89  90  91  92  93  95  96  97  98  99 101 102 103 105 106 107 108 109 110 111 112
  4   9   8   4   7  15   5   3  17   7  16   9  10   6   4  10   3   2   1   1   2   2   1   1   1   1   1   9   2   1   1  13   2   3  15   2
113 114 115 117 118 119 120 121 122 123 124 125 126 128 130 131 133 134 135 136 139 140 143 144 145 147 148 149 150 151 152 153 155 156 164 166
  5   7   5   8   7   8   3   2   2   1   9   4   3   7   4   1   5   3   1   1  28   1   4   1   2   3   7   1   3   4   3   2   2   1   1   1
169 170 172 175 178 195 207 209 222 228 233 246 248 253 260
  1   1   2   1   1   1   2   1   1   1   2   1   2   2   1
```

Figure 6: *Summary of Pixel Rate*

3.4.4 Texture Rate

The texture rate data has a widespread distribution similar to pixel rates, with a large concentration of GPUs possessing moderate texture processing capabilities.

```
Texture_Rate
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
  1  56 135  53 102 116  89  30  39  43  27  46  22  59 110  17  68  37  36  47  24  19  18   9  28  49  52  19  25  24  18  16  17   8  52  18  21
 37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73
 25  19  11  20  13   7  17  20  26  12  11  20   4  25  21  16  14   3  14  16   7  40  28  25  19  43   6  14   8  12  35  20  10  18   3   9   3
 74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  99 100 101 102 103 104 105 106 107 108 109 110 111 112
  7  12  10   6   9  15  18   9  22  20  23  15  18   3  19   6   8   6   9   2   3   1   4  10   2  12   4   8   9   4   3   5   9  12  32   6   6
113 114 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150
 11   4   2   5  15   2   1  12   6  18   5   5  10  10  15   6   4   3   8   4   9  10   6  16   5   8   4  18  11   4   7   1   5   8   5   6   2
151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 172 173 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189
  9   8   1  13   2  22   2   1  12  10  11   9   2   7   2   7   3   4   7   5   5   5   2   7   2   6   4   1   4  26   8   2  11   4   4   5   9
190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 213 214 215 216 217 218 219 220 223 224 225 227 228 230 232
  3   3   3  14   4   2  10   9   5   5   2   5  15   4   1   4   4   5   2   8   2   1   2   3   1   7   1   4   1   7   6   2   4   8   1   4   4
233 235 237 238 240 241 243 245 246 249 251 253 254 256 257 260 261 266 268 269 271 273 274 276 277 278 279 282 283 284 285 287 288 292 294 296 298
  1   1   1   1   1   1   1   2   1   1   4   1   1   4   3   4   3   1   1   3   1   1   3   2  15   1   2   2   2   3   1   1   1   1   1   5   2
303 304 306 308 310 311 319 320 327 328 329 333 337 340 343 344 346 350 352 354 357 358 363 365 366 368 370 374 377 379 381 383 384 386 392 394 396
  1   1   1   2   1   1   1   5   1   1   1   2   1   1   1   2   2   1   1  30   2   2   3   1   4   2   2   2   5   2   1   1   2   4   2   1   1
410 418 419 420 439 445 450 452 454 464 467 477 506 512 522 538 542 555 717
  1   1   1   2   1   1   1   1   2   1   2   1   1   1   1   1   1   1   1
```

Figure 7: *Summary of Texture Rate*

3.4.5 PSU

The dataset showcases a broad range of power supply requirements, with many units clustered around 300-500 Watt and 20-30 Amps, indicating a standard for mainstream hardware.

```
PSU
         1000 Watt 1000 Watt & 42 Amps 1000 Watt & 46 Amps 1000 Watt & 50 Amps 1000 Watt & 67 Amps           1250 Watt 1500 Watt & 50 Amps
                 1                   3                   1                   1                   2                   2                   1
 250 Watt & 17 Amps  250 Watt & 18 Amps  30 Watt & 127 Amps            300 Watt  300 Watt & 12 Amps  300 Watt & 15 Amps  300 Watt & 16 Amps
                 1                   5                   1                 201                   2                   2                   2
 300 Watt & 18 Amps  300 Watt & 20 Amps  300 Watt & 22 Amps  300 Watt & 27 Amps            350 Watt  350 Watt & 17 Amps  350 Watt & 18 Amps
                74                 114                 182                  59                 202                   2                  10
 350 Watt & 20 Amps  350 Watt & 22 Amps  350 Watt & 23 Amps  350 Watt & 24 Amps  350 Watt & 26 Amps            400 Watt  400 Watt & 18 Amps
                 6                   5                  61                   8                  26                 152                  26
 400 Watt & 20 Amps  400 Watt & 22 Amps  400 Watt & 23 Amps  400 Watt & 24 Amps  400 Watt & 25 Amps  400 Watt & 26 Amps  400 Watt & 27 Amps
                82                   6                   2                  10                  15                  17                 121
 400 Watt & 28 Amps  400 Watt & 29 Amps  400 Watt & 30 Amps  400 Watt & 33 Amps  400 Watt & 37 Amps  400 Watt & 42 Amps  420 Watt & 20 Amps
                 4                 152                   1                   2                   3                   1                   2
 420 Watt & 30 Amps            450 Watt  450 Watt & 18 Amps  450 Watt & 20 Amps  450 Watt & 23 Amps  450 Watt & 24 Amps  450 Watt & 26 Amps
                 1                 292                   1                   9                   1                  31                   4
 450 Watt & 28 Amps  450 Watt & 29 Amps  450 Watt & 30 Amps  450 Watt & 32 Amps  450 Watt & 33 Amps  450 Watt & 34 Amps  450 Watt & 38 Amps
                 7                   1                 138                   1                  37                   3                   4
          500 Watt  500 Watt & 20 Amps  500 Watt & 24 Amps  500 Watt & 25 Amps  500 Watt & 26 Amps  500 Watt & 27 Amps  500 Watt & 28 Amps
               170                   4                   6                   6                   2                   1                   3
 500 Watt & 30 Amps  500 Watt & 31 Amps  500 Watt & 32 Amps  500 Watt & 33 Amps  500 Watt & 34 Amps  500 Watt & 35 Amps  500 Watt & 37 Amps
                80                   2                   4                 263                   1                  10                  44
 500 Watt & 38 Amps  500 Watt & 40 Amps  500 Watt & 42 Amps  520 Watt & 25 Amps            550 Watt  550 Watt & 20 Amps  550 Watt & 28 Amps
                 2                  13                  36                   1                  10                   2                   1
 550 Watt & 30 Amps  550 Watt & 31 Amps  550 Watt & 33 Amps  550 Watt & 34 Amps  550 Watt & 35 Amps  550 Watt & 37 Amps  550 Watt & 38 Amps
                 8                   2                   1                   5                   3                   8                  27
 550 Watt & 39 Amps  550 Watt & 40 Amps  550 Watt & 42 Amps  550 Watt & 46 Amps  580 Watt & 35 Amps            600 Watt  600 Watt & 26 Amps
                 1                   8                   4                   5                   1                  38                   1
 600 Watt & 29 Amps  600 Watt & 30 Amps  600 Watt & 38 Amps  600 Watt & 40 Amps  600 Watt & 42 Amps  600 Watt & 50 Amps  600 Watt & 60 Amps
                 1                   4                   1                   1                 226                   4                   2
          650 Watt  650 Watt & 24 Amps  650 Watt & 28 Amps  650 Watt & 30 Amps  650 Watt & 34 Amps  650 Watt & 38 Amps  650 Watt & 40 Amps
                 8                   3                   2                   1                   1                   2                   2
 650 Watt & 42 Amps  650 Watt & 43 Amps  650 Watt & 50 Amps  680 Watt & 46 Amps  680 Watt & 50 Amps  700 Watt & 30 Amps  700 Watt & 40 Amps
                 4                   2                   2                   3                   1                   1                   1
 700 Watt & 42 Amps  700 Watt & 46 Amps  700 Watt & 50 Amps  700 Watt & 51 Amps            750 Watt  750 Watt & 30 Amps  750 Watt & 32 Amps
                 6                   5                   1                   1                 153                   3                   1
 750 Watt & 35 Amps  750 Watt & 38 Amps  750 Watt & 40 Amps  750 Watt & 42 Amps  750 Watt & 45 Amps  750 Watt & 46 Amps  750 Watt & 50 Amps
                 1                   1                   4                  31                   2                   1                  26
 750 Watt & 55 Amps  800 Watt & 50 Amps  800 Watt & 53 Amps  850 Watt & 42 Amps  850 Watt & 50 Amps  850 Watt & 56 Amps  850 Watt & 62 Amps
                 4                   6                   3                   1                   5                   1                   1
 850 Watt & 71 Amps  900 Watt & 42 Amps  900 Watt & 60 Amps  950 Watt & 79 Amps
                 1                   1                   1                   1
```

Figure 8: *Summary of PSU*

3.4.6 TMUs

TMUs also display a broad distribution, with a focus on GPUs possessing between 32 to 128 TMUs, suggesting these are common configurations in the dataset.

```
TMUs
  1   2   4   8  12  14  16  20  24  28  32  36  40  44  48  56  60  64  72  80  88  96 104 106 112 120 128 144 160 176 192 224 240 256 320 384
 41   7 157 392   9   1 280  92 177   6 294  12 208   6 131 186  21 297  79 166  12  67  59   1 147  38 221  44  62  88  39  29  24   9   2   2
```

Figure 9: *Summary of TMUs*

3.4.7 ROPs

The ROPs count for most units is on the lower end, with many GPUs featuring between 16 to 64 ROPs. A small number have very high ROP counts, likely representing high-end units.

```
ROPs
      1      12     128 14 (x2)      16 16 (x2) 16 (x4)       2      20      22 22 (x4)      24 24 (x2) 24 (x3) 24 (x4)      28 28 (x2)       3
     59       9       3       1     641     109       4      58       3       1      29     106      17       1       1       9       6       4
     32 32 (x2) 32 (x3) 32 (x4)       4  4 (x4)      40 40 (x2) 44 (x2)      48 48 (x2) 48 (x3) 48 (x4)      56 56 (x2)       6      64 64 (x2)
    736      72       2       1     568       1      19       2       3     135      20       1       2      46      12       2     187      30
64 (x3)       8  8 (x2)      88      96 96 (x2)
      2     432       4      14      44      10
```

Figure 10: *Summary of ROPs*

4 Descriptive statistics

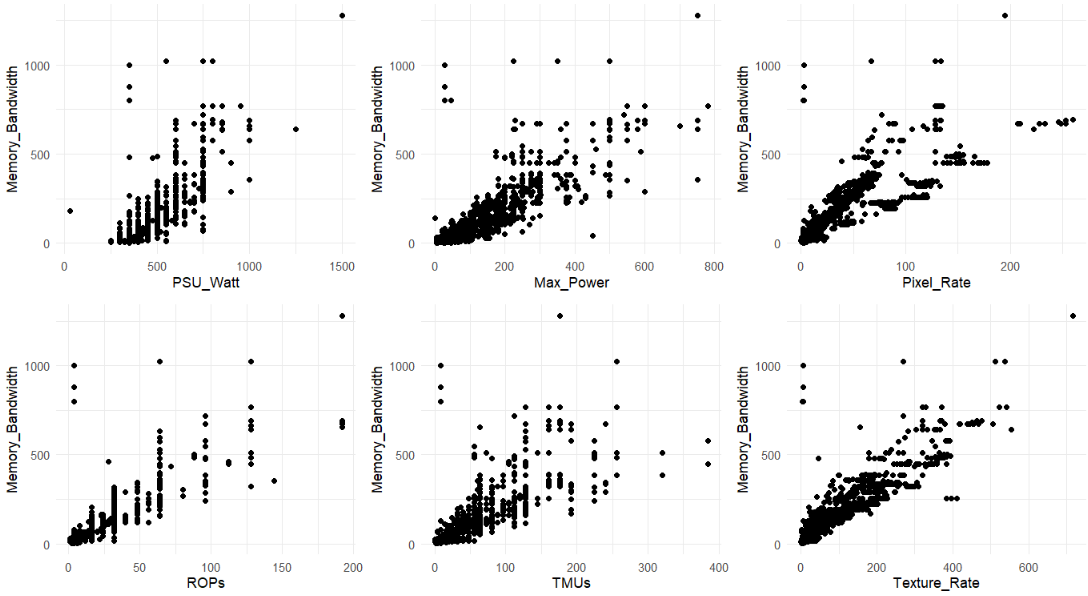


Figure 11: *Manufacturers plot*

These six scatter plots in Figure 11 shows correlation between Memory Bandwidth with PSU, Max Power, Pixel Rate, ROPs, Texture Rate and TMUs in relation to non-voluntary context switches in computing systems. Scatter plots are graphical representations where each dot represents an observation [12]. The position of a dot on the horizontal and vertical axis indicates values for an individual data point. The following encapsulates the essence of each plot, with a more comprehensive examination to be conducted in Section 5.3.2

- **1st plot:** The x-axis is labeled "PSU Watt" and goes from 0 to 1500. The y-axis is labeled "Memory Bandwidth" and goes from 0 to 1000.
- **2nd plot:** The x-axis is labeled "Max Power" and goes from 0 to 800. The y-axis is labeled "Memory Bandwidth" and goes from 0 to 1000.
- **3rd plot:** The x-axis is labeled "Pixel Rate" and goes from 0 to 350. The y-axis is labeled "Memory Bandwidth" and goes from 0 to 1000.
- **4th plot:** The x-axis is labeled "ROPs" and goes from 0 to 200. The y-axis is labeled "Memory Bandwidth" and goes from 0 to 1000.
- **5th plot:** The x-axis is labeled "TMUs" and goes from 0 to 400. The y-axis is labeled "Memory Bandwidth" and goes from 0 to 1000.
- **6th plot:** The x-axis is labeled "Texture Rate" and goes from 0 to 600. The y-axis is labeled "Memory Bandwidth" and goes from 0 to 1000.

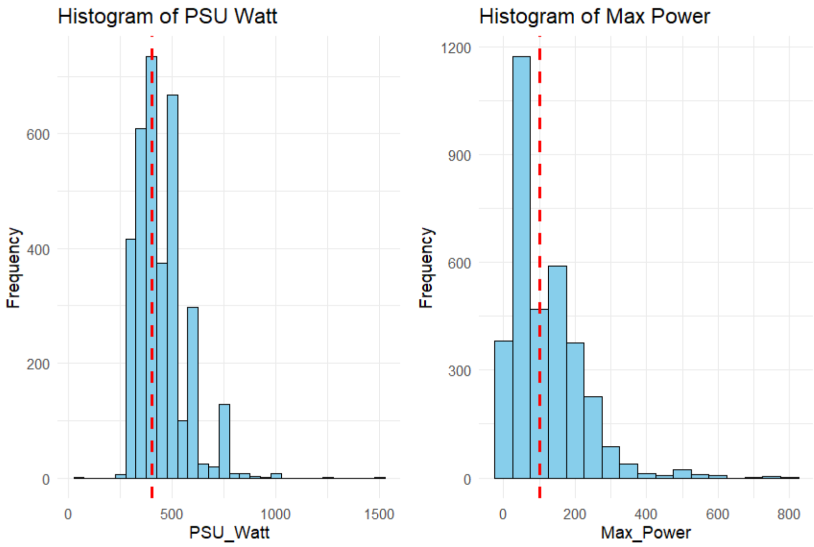


Figure 12: *Histogram of PSU Watt and Max Power*

These histogram of PSU Watt and Max Power in Figure 12 are graphical representations that organize a group of data points into a specified range. [3]

Here are some things we can see from this histogram:

- **PSU Watt Histogram:**The first histogram on the left represents the distribution of PSU (Power Supply Unit) Watt. It appears that most PSUs in this dataset have a wattage around 500W. This means that 500W is the most common power output for the PSUs in our dataset.
- **Max Power Histogram:**The second histogram on the right shows the distribution of Max Power. The majority of max power values are around 100W. This indicates that 100W is the most frequent maximum power output in our dataset.
- The red line in each histogram represents for the median. The median of PSU Watt is around 480W and Max Power is 100W

The red line in each histogram is the median. The median of PSU Watt is around 480W and Max Power is 100W

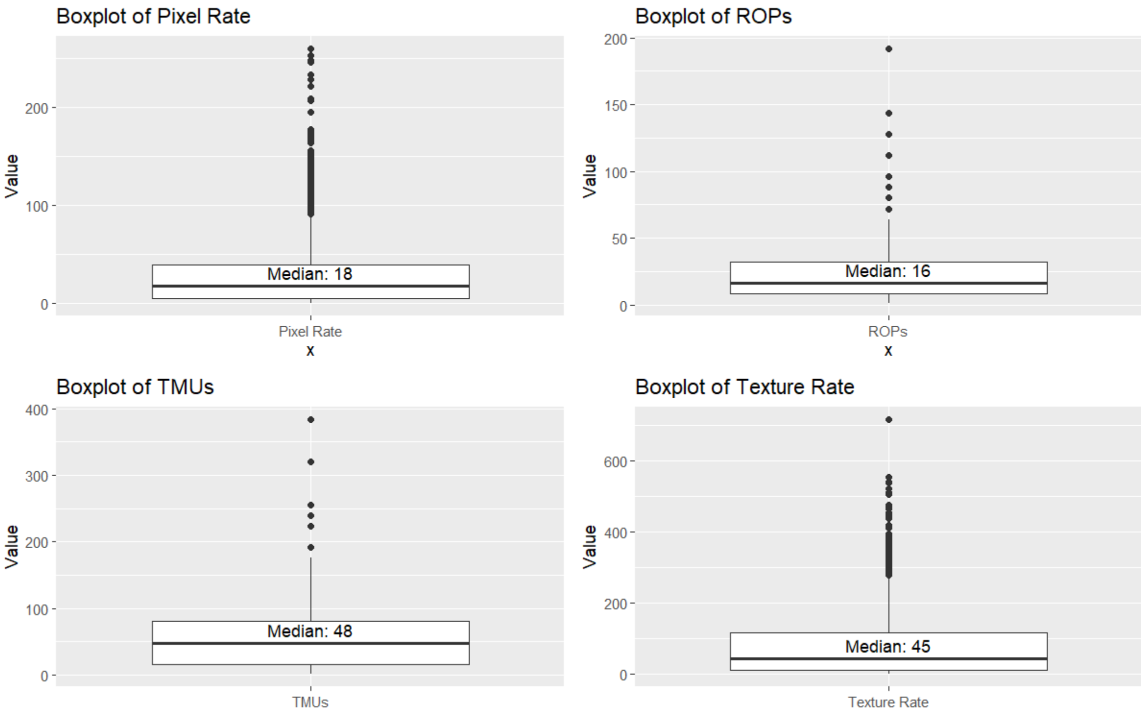


Figure 13: *Box plot*

In Figure 13, these box plots [11] represent the distribution of different variables such as Pixel Rate, ROPs, TMUs, and Texture Rate.

- **Pixel Rate:** The box plot is compact, indicating that the data points are closely packed with a small interquartile range. There are several outliers above the upper whisker, showing individual data points that fall far from the main group. The line inside the box indicates the median value, which is the middle value of the dataset. For this plot, the median value is 18
- **ROPs:** The box is quite thin, it seems like most of the ROPs values are relatively low, but there are some higher values as well. Multiple outliers are scattered vertically above the upper whisker. The line inside the box indicates the median value, which is the middle value of the dataset. For this plot, the median value is 16.
- **TMUs:** The box in the boxplot represents the interquartile range. The line inside the box represents the median of the data, which is the 50th percentile. In this case, the median value is 48. It just has a few outliers extended to around 400
- **Texture Rate:** It features a moderate interquartile range and visible outliers. The median line is closer to the bottom of the box and its value is 45. The points above the upper whisker are considered outliers, in this case, there are several outliers extending up to a value close to 600.

5 Inferential statistic

5.1 Target

Our objective in this report is to forecast the memory bandwidth value using pertinent variables extracted from the original dataset. To achieve this, we will employ **multiple linear regression**. Our method involves dividing the dataset into two equal halves: one for training the model and the other for testing.

5.2 Training

5.2.1 Model definition

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \epsilon$$

where

- y is `Memory_Bandwidth`
- $x_1 ... x_7$ are `PSU_Amps`, `PSU_Watt`, `Max_Power`, `Pixel_Rate`, `ROPs`, `TMUs`, `Texture_Rate`
- β_0 is the intercept
- β_i, $i = 1...7$ are coefficients of the independent variables
- ϵ : error term, the function `lm()` in R code will automatically accounted when fitting the model

5.2.2 Model fitting

Using `lm` function, we perform hypothesis test for each predictor:

$$H_0 : \beta_i = 0, i = 1...7$$
$$H_1 : \beta_i \neq 0, i = 1...7$$

- **Null Hypothesis** H_0: indicates there is no relationship between `Memory_Bandwidth` and other variables
- **Alternative Hypothesis** H_1: indicates there is a relationship between `Memory_Bandwidth` and other variables
- **t-value**: the higher the t-value, the greater the confidence we have in the coefficient as a predictor
- **p-value**: the probability of observing the t-statistic, given that the null hypothesis is true. Typically, p-value less than 0.05 means statistical significance.

5.2.3 Result

```
Call:
lm(formula = Memory_Bandwidth ~ Max_Power + PSU_Watt + PSU_Amps +
    Pixel_Rate + ROPs + TMUs + Texture_Rate, data = traindata)

Residuals:
    Min      1Q  Median      3Q     Max
-304.37  -14.04   -3.04    8.24  982.11

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -29.74195    9.49604  -3.132  0.00177 **
Max_Power      0.33851    0.03548   9.541  < 2e-16 ***
PSU_Watt       0.01576    0.02407   0.655  0.51274
PSU_Amps       0.88845    0.32115   2.766  0.00573 **
Pixel_Rate    -0.49635    0.29317  -1.693  0.09063 .
ROPs           1.73877    0.29274   5.940 3.47e-09 ***
TMUs          -0.03677    0.05614  -0.655  0.51255
Texture_Rate   0.87342    0.05583  15.643  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.17 on 1664 degrees of freedom
Multiple R-squared:  0.8483,Adjusted R-squared:  0.8477
F-statistic:  1330 on 7 and 1664 DF,  p-value: < 2.2e-16
```

Figure 14: *Output of* `summary(train_model)`

From Figure 14, we receive some notable points:

- **Significance levels**: there are 3 predictors such as `PSU_Watt`, `Pixel_Rate`, and `TMUs` have p-value greater than 0.05, which means these variables are not statistical significant
- **Adjusted R - squared**: tells us how well the independent variables collectively explain the variation in the dependent variable. In this case, this value is approximately 84.77% suggests that the model is able to capture a large proportion of the variability in the dependent variable
- **F-statistic:** provides valuable information about the overall fit of the regression model. In this case, it is equal to 1330 with an extremely low p-value indicates that the F-statistic is highly significant

Therefore, we can reject the null hypothesis, which assumes that all regression coefficients are equal to zero, and conclude that the regression model is statistically significant overall.

5.2.4 Checking assumption

There are some assumptions needed to be checked in our model:

- First, the residual values are normally distributed.

- Second, we need to make sure that there must be a linear relationship between the dependent and the independent variables. This could be seen through these scatter plots:
- Third, we have to make sure that the independent variables are not highly correlated with each other or we can call it multi-collinearity. This could be verified by computing a mattrix of correlaton coefficients among the independent variables. Thus, all the coefficients should be less than 0.8.
- Finally, we need to make sure that the variance of the residual errors is similar across the value of each independent variable.

5.2.4.a Normality

This assumption can be checked straight forward through the Q-Q plot - a graphical technique for assessing if a dataset is normally distributed.

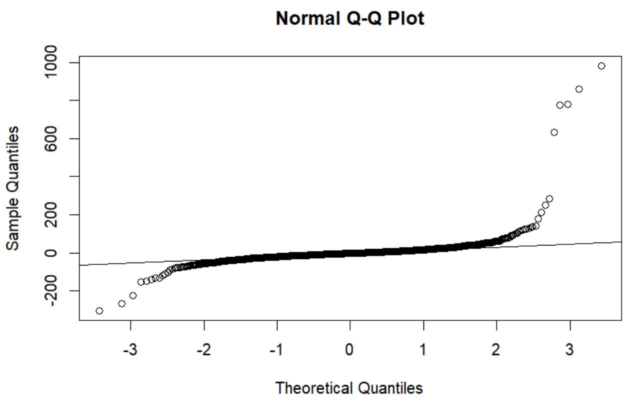


Figure 15: *Q-Q plot of residuals extract from dataset*

In Figure 15, the plot shows that most data points align well with the reference line in the central part of the plot, indicating normality in the middle range of data. At the both ends, the data points deviate from the reference line, suggesting that the distribution has heavier tails than a normal distribution. This pattern may imply skewness in the data, with a possible right skew due to the upward deviation on the right end.

For clearer evaluate, we consider the histogram of the residual:

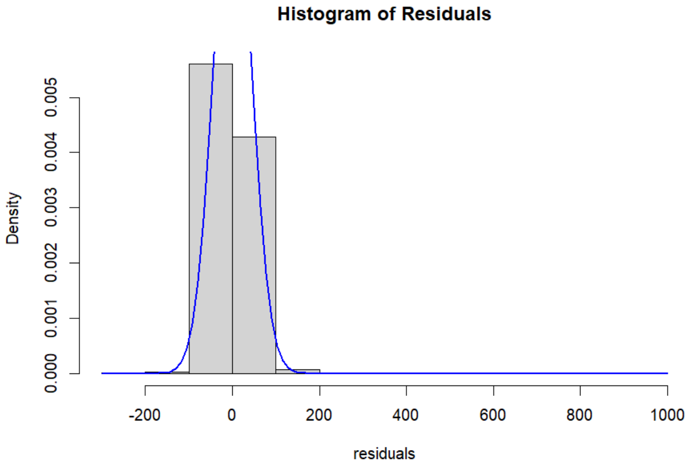


Figure 16: *Histogram of residuals extract from dataset*

In Figure 16, the histogram of residuals shows that the residuals are not symmetrically distributed, with a concentration of values around zero. A very small portion of data has the values greater than 500, also, there are few data points seperately place far from the remain data (which is near the reference line). Those points may be confounding values. After remove the data points of residuals that higher than 500, we got the graph:

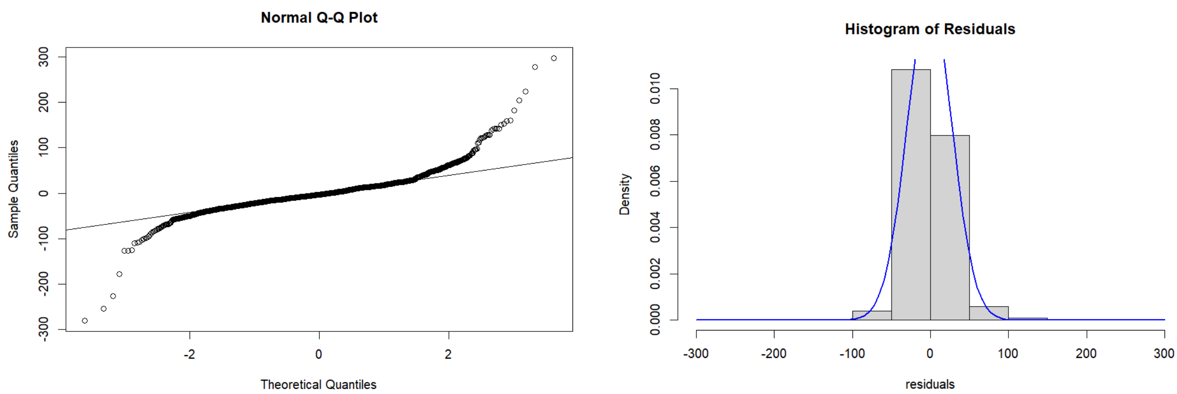


Figure 17: *Histogram and Q-Q plot of residuals extract from dataset*

In Figure 17, the histogram indicates a slight skewness as the bars are not symmetrically distributed around the center but still perform a form of normality distribution.

When considering both the Q-Q plot and the histogram, it appears that the dataset does not fully meet the assumption of normality. There are indications of skewness and potential outliers affecting the distribution.

5.2.4.b Linear relationship

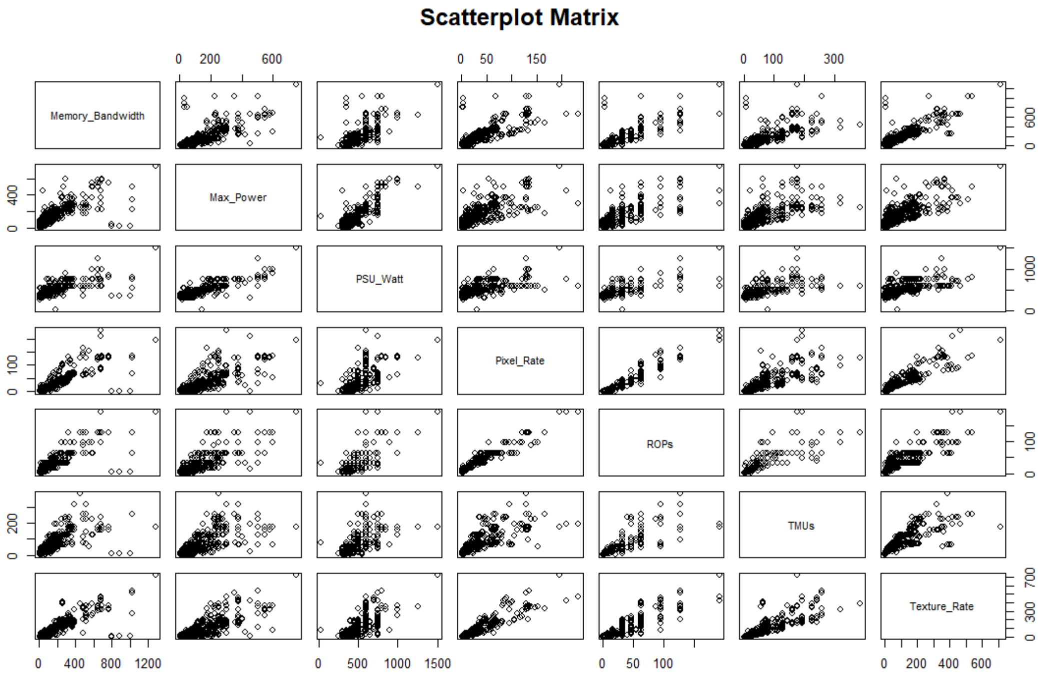


Figure 18: *Scatterplot Matrix*

Based on the Figure 18, it seems like there are various relationships between Memory Bandwidth and different features in computing systems. Let's break down the linear relationships between Memory Bandwidth and each of the specified parameters: PSU Watt, Max Power, Pixel Rate, ROPs, TMUs, and Texture Rate.

- **PSU Watt**: The scatter plot shows a general upward trend, indicating a positive correlation between Memory Bandwidth and PSU Watt. As PSU Watt increases, Memory Bandwidth tends to increase as well. This suggests that higher power consumption by the system is associated with greater memory bandwidth.
- **Max Power**: As Max Power increases, the data points start to disperse, suggesting a wider range of Memory Bandwidth. While there might not be a clear linear relationship, the dispersion of data points could indicate that higher Max Power is associated with a greater variability in Memory Bandwidth.
- **Pixel Rate**: There is a clear positive correlation between Pixel Rate and Memory Bandwidth. Higher pixel rates require more data to be transferred, leading to higher Memory Bandwidth. This relationship aligns with the intuitive understanding that higher pixel rates necessitate more data processing and transfer.
- **ROPs (Raster Operations Pipelines)**: Each data point represents a measurement of ROPs for a specific memory bandwidth. While the description doesn't explicitly mention the nature of the relationship, it's likely that there is a positive correlation. Higher ROPs often require more memory bandwidth to support the rendering of graphics and images.

- **TMUs (Texture Mapping Units)**: The sparse data points at higher values suggest that fewer configurations have high values for both TMUs and Memory Bandwidth. This indicates a relationship where higher TMUs might not necessarily lead to a proportional increase in Memory Bandwidth. However, without more details, it's challenging to ascertain the exact nature of this relationship.
- **Texture Rate**: The dispersion of data points as Texture Rate increases could indicate varying performance characteristics. This suggests that as Texture Rate increases, Memory Bandwidth can vary significantly. There might not be a straightforward linear relationship between Texture Rate and Memory Bandwidth, as other factors could influence performance.

In summary, while some relationships like Pixel Rate and Texture Rate show a clear positive correlation with Memory Bandwidth, others like ROPs and PSU Watt might have more complex or variable relationships.

5.2.4.c Multicolinearity

Multicollinearity refers to the phenomenon where independent variables in a regression model are highly correlated. This condition can significantly impair the model's ability to provide reliable and interpretable statistical estimates. It affects the precision of the estimate coefficients, leading to unreliable statistical inferences. This section of the report focuses on examining the presence of multicollinearity among the variables in our model and discusses its implications.

5.2.4.1 Data and Methods: The analysis was conducted using a multiple linear regression model where `Memory_Bandwidth` was modeled as a function of several predictors: `Max_Power`, `PSU_Watt`, `Pixel_Rate`, `ROPs`, `TMUs`, and `Texture_Rate`. To assess multicollinearity, we examined scatterplot matrices, residuals versus fitted plots, and normal Q-Q plots.

5.2.4.2 Findings

1. **Scatterplot Matrix Analysis:**

   The scatterplot matrix provided a visual assessment of the relationships between the predictors. Notably, strong linear relationships were observed between `Pixel_Rate`, `ROPs`, and `Texture_Rate`. Such relationships suggest potential multicollinearity issues as these variables exhibit strong intercorrelations which can obscure the individual effect of each predictor on the dependent variable.

2. **Residuals vs. Fitted Values Plot**

   The residuals vs. fitted values plot was used to check for non-random error patterns. The plot exhibited a fanning out of residuals at higher fitted values, suggesting heteroscedasticity, which often accompanies multicollinearity. This heterogeneity in variance could be indicative of an underlying issue with correlated predictors affecting the model's assumptions.

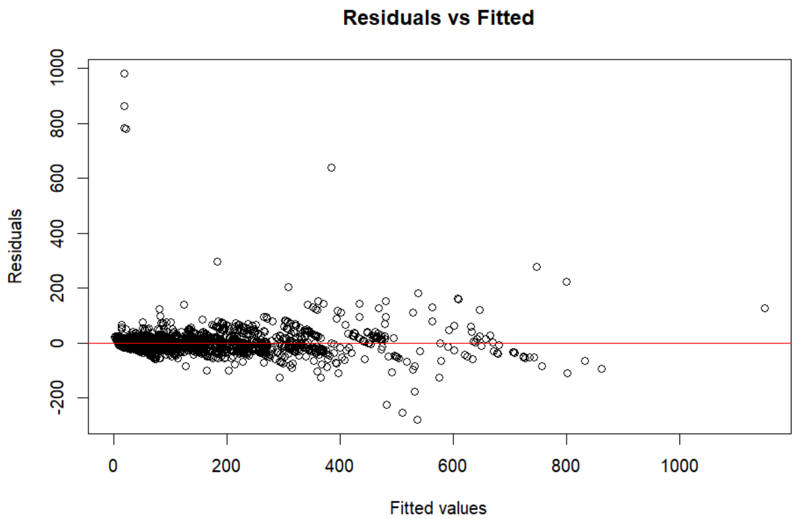


Figure 19: *Residuals vs Fitted*

Description of the Residuals vs Fitted Graph in Figure 19:

- Horizontal Axis (Fitted Values): Represents the predicted values from the regression model, denoted as $\hat{y}$.
- Vertical Axis (Residuals): Represents the residuals for each observation, calculated as the difference between the actual and the predicted values $(y - \hat{y})$.
- Red Line: A horizontal line at zero residual value, helping to highlight the deviations of the residual points from zero.

Detailed Analysis:

Residual Distribution:

- The residuals do not appear uniform as the predicted values increase. Notably, the residuals seem to exhibit greater variability at higher ranges of the predicted values.
- The data points are not evenly distributed around the zero line, indicating potential issues with homoscedasticity (constant variance).

Residual Patterns:

- The slope seen on the right side of the graph may indicate that the model does not fit well at higher prediction ranges. This could be a sign that the model fails to capture all the factors affecting the response, or it could indicate model overfitting.

Signs of Outliers or High Leverage Points:

- Some points far from the zero line may be outliers or points with high influence, especially those at the far right and left ends of the graph.

3. **Normal Q-Q Plot:**

The normal Q-Q plot was assessed to check the normality of residuals. In Figure 15, the deviation of points from the line in the tails indicates potential outliers or skewness due to influential points, which could be related to issues in the predictor variables, including multicollinearity.

The analysis clearly points to multicollinearity particularly between `Pixel_Rate`, `ROPs`, and `Texture_Rate`. The inter-correlation among these variables can lead to inflated variance inflation factors (VIFs), reduced power of the model, and coefficients that may be poorly estimated and highly sensitive to changes in the model. Such conditions compromise the reliability and interpret ability of the regression model.

Conclusion:

The presence of multicollinearity in our regression model particularly among `Pixel_Rate`, `ROPs`, and `Texture_Rate` poses challenges to deriving clear and reliable insights. It is imperative to address these issues to improve the model's performance and ensure the robustness of our conclusions.

5.2.4.d Homoscedasticity

In this section, we will try to examine the homoscedasticity of the training model. Thus, we can use the plot of residual against fitted value below, if there is a pattern or a funnel shape in the plot, then it is heteroscedastic scenario, if it is a homoscedastic scenario, then the spread of the residuals should be roughly constant across the range of fitted values.

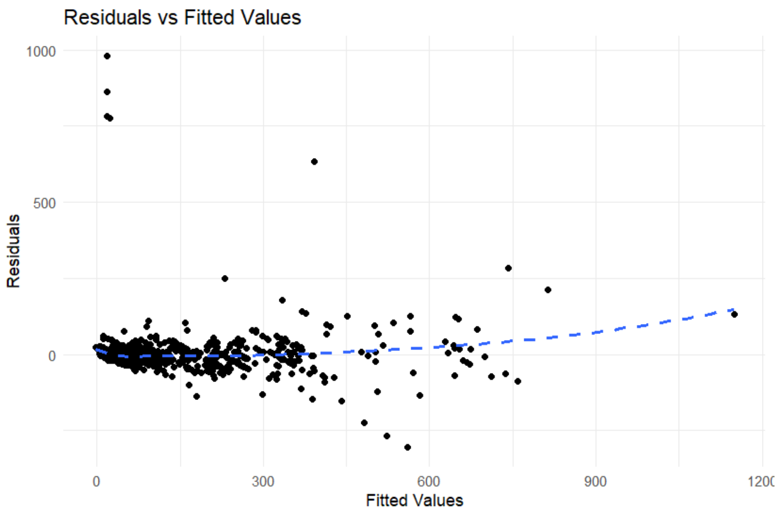


Figure 20: *Pattern of the relationship between residual and fitted value*

In Figure 20, we could see the pattern of the relationship between the residual and the fitted value is increasing. So it might suggest heteroscedasticity (non-constant variance).

In addition, we have used Breusch-Pagan Test to complete our examination.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 106.9069, Df = 1, p = < 2.22e-16
```

Figure 21: *Breusch-Pagan Test*

In Figure 21, as we can we that the p-value is less than or equal to 2.22e-16, which is less than the significant level(alpha) usually set at 0.05. So we reject the null hypothesis (constant variance) in the residuals of the linear regression model.

5.3 Testing

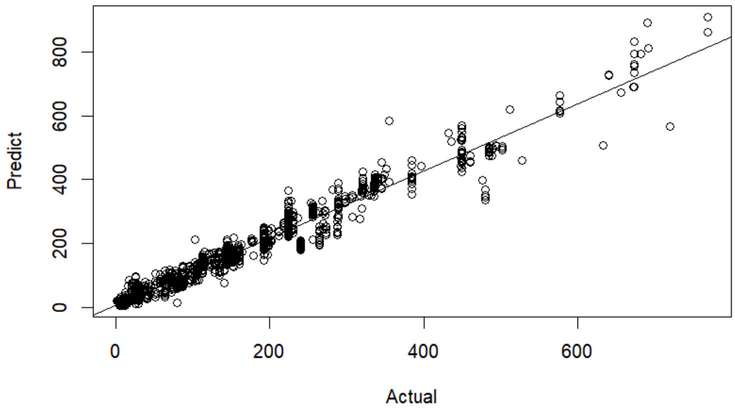


Figure 22: *Scatter plot of predicted and actual value of* `Memory_Bandwidth`

In Figure 22, the regression line closely mirrors the diagonal, suggesting that the model effectively predicts the majority of values. However, there are notable outliers, particularly at higher values (> 600), hinting that the model's efficacy diminishes when dealing with instances of significant magnitude. This observation underscores the need for further analysis to understand the underlying factors contributing to these outliers and potentially refine the model to improve its predictive performance, especially in scenarios involving larger values.

6 Discussion and extension

6.1 Multiple Linear Regression

6.1.1 Advantages

These are some advantages of Multiple Linear Regression [13]:

- It offers a high degree of interpretability. Each coefficient in an MLR model represents the change in the response variable for each one-unit change in the corresponding explanatory variable, assuming all other variables are held constant. This makes MLR a highly interpretable model.
- MLR can be a powerful predictive tool. If the model's assumptions are met, the model can be used to predict the response variable with a high degree of accuracy.
- MLR allows for the control of confounding variables, which are variables that are correlated with both the explanatory and response variables.By including these confounding variables in the model, we can estimate the effect of the explanatory variables on the response variable more accurately.

6.1.2 Disadvantages:

Multiple Linear Regression also has its disadvantages [13]:

- One of the main disadvantages is that MLR makes several key assumptions, including linearity, independence, homoscedasticity, and normality. If these assumptions are violated, then the model's predictions can be unreliable.
- Another disadvantage is multicollinearity, which occurs when two or more explanatory variables are highly correlated with each other. This can make it difficult to determine the effect of each variable independently and can also lead to unstable estimates of the regression coefficients, which can make the model's predictions unreliable.
- Lastly, if a model includes too many explanatory variables, particularly variables that are not relevant to the response variable, the model can become over fit. An over fit model will perform well on the training data but poorly on new, unseen data. This is because the model has fit to the noise in the training data, rather than the underlying trend.

6.2 Extension using Ridge Regression

6.2.1 Advantages

- **Reduction of Multicollinearity**: Ridge Regression effectively addresses the issue of multicollinearity in data by adding a penalty to the size of the coefficients. This regularization technique helps stabilize the coefficients, even when independent variables are highly correlated, thus enhancing the generalization of the model.
- **Shrinkage of Coefficients**: Ridge introduces a shrinkage penalty ($\lambda \sum_{j=1}^{p} \beta_j^2$) that controls the magnitude of the regression coefficients. This prevents any single predictor from exerting too much influence on the model, which is particularly beneficial in models with many predictors.

- **Improved Model Stability**: By penalizing large coefficients, Ridge helps to reduce the model's sensitivity to noise in the data. This results in a more stable and reliable model, which is less likely to overfit compared to ordinary least squares regression.
- **Bias-Variance Tradeoff**: Ridge Regression manages the bias-variance tradeoff effectively. Although it introduces a slight bias into the estimates by shrinking the coefficients, it significantly reduces the variance of the model predictions, often leading to better long-term prediction performance.
- **Computational Efficiency**: Despite the introduction of a penalty term, Ridge Regression can be efficiently computed using matrix operations, making it computationally scalable to handle large datasets..

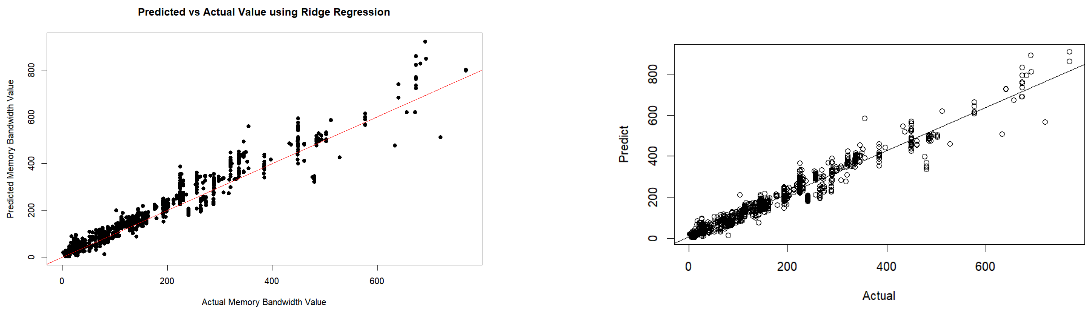
6.2.2 Disadvantages:

- **Bias Introduction**: One of the main drawbacks of Ridge Regression is that it introduces bias into the estimates through the shrinkage of coefficients. This bias can sometimes lead to underfitting if the penalty term λ is set too high.
- **Variable Selection**: Unlike Lasso Regression, which can zero out coefficients for some predictors, Ridge Regression does not inherently perform variable selection. All predictors included in the initial model will remain in the model with their coefficients shrunk towards zero, but not exactly zero. This can make model interpretation more challenging, especially when dealing with high-dimensional data.
- **Parameter Sensitivity**: The performance of Ridge Regression heavily depends on the choice of the penalty parameter λ. Selecting an appropriate value of λ is crucial and can be challenging. It typically requires cross-validation or other tuning techniques, which can be computationally intensive.
- **Lack of Sparsity**: Ridge Regression does not produce sparse models and hence may not be the best choice when the goal is to identify a reduced set of predictors that have substantial influence on the response variable.
- **Assumption of Linearity**: Like other forms of linear regression, Ridge assumes a linear relationship between the predictors and the response variable. This assumption may not hold in scenarios where the underlying relationships are non-linear, thus limiting the applicability of Ridge Regression in such cases.

6.2.3 Comparison of Ridge Regression and Multiple Linear Regression

6.2.3.a Visual Analysis:

1. **Ridge Regression:**
   - The first graph depicts the results from a Ridge Regression model. The plot shows a strong linear relationship between the actual and predicted values. The points cluster closely around the red line, which represents the ideal where predictions perfectly match actual values. Notably, there is a distinct pattern where higher values tend to deviate slightly above the line, suggesting slight over-predictions in the higher range.
2. **Multiple Linear Regression:**



(a) *Ridge Regression* **(b)** *Multiple Linear Regression*

- The second graph presents the output from a Multiple Linear Regression model. Similar to the Ridge Regression, there is a clear linear relationship between the predicted and actual values. The points also adhere closely to the diagonal line indicating good model performance across most of the data range. However, compared to the Ridge Regression, the spread of points around the diagonal line appears slightly wider, especially in the mid-range of values.

6.2.3.b Model Performance:

- **Consistency Across Data Range:**
  - **Ridge Regression** tends to show better consistency in predictions across the entire range of actual values, especially by providing slightly tighter clusters of predicted values around the line, which may indicate better handling of multicollinearity or outliers within the data.
  - **Multiple Linear Regression**, while effective, shows a bit more variability in prediction accuracy, especially in the central range of the data. This variability could be due to the model's sensitivity to high collinearity among predictors, which is less effectively managed than in Ridge Regression.
- **Outlier Influence:**
  - **Ridge Regression** appears more robust against outliers, as indicated by the uniformity in the spread of residuals. This robustness helps in improving the model's generalizability and reducing the error variance.
  - **Multiple Linear Regression** might be slightly more affected by outliers or extreme values, as evidenced by the broader spread of points, which could lead to higher variability in predictions.

6.2.3.c Conclusion:

- Both models perform commendably, indicating strong predictive capabilities. However, Ridge Regression may offer a slight advantage in scenarios where predictor variables exhibit high multicollinearity, as it effectively reduces the impact of this collinearity on model predictions.
- The choice between using Ridge Regression and Multiple Linear Regression should consider the specific characteristics of the data set, including the presence of multicollinearity and

the range of data values to be predicted. Ridge Regression is particularly useful when the data includes highly correlated predictors, while MLR could be preferable for simpler, less collinear datasets or when interpretability is a critical factor.

7 Data and code availability

- Link data: data
- Link code: code

References

[1] analyticsvidhya. *Best way to learn kNN Algorithm using R Programming*. https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/. Accessed: 2024-05-05.

[2] David R. Foxcroft Danielle J. Navarro. *Checking the normality of a sample*. https://lsj.readthedocs.io/ko/latest/Ch11/Ch11_tTest_08.html. Accessed: 2024-04-04.

[3] Geeks for Geeks. *Density Plots in R*. https://www.geeksforgeeks.org/histograms-and-density-plots-in-r/. Accessed: 2024-04-20.

[4] Geeks for Geeks. *How to Find The Optimal Value of K in KNN* . https://www.geeksforgeeks.org/how-to-find-the-optimal-value-of-k-in-knn/. Accessed: 2024-05-05.

[5] GamersNexus. *GPU Dictionary: Understanding GPU Video Card Specs*. https://gamersnexus.net/guides/717-gpu-dictionary-understanding-gpu-video-card-specs. Accessed: 2024-05-05.

[6] Intel. *What Is a GPU? Graphics Processing Units Defined*. https://www.intel.com/content/www/us/en/products/docs/processors/what-is-a-gpu.html. Accessed: 2024-04-02.

[7] Mark Elliot Maria Pampaka Mark Tranmer, Jen Murphy. *Multiple Linear Regression*. https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/multiple-linear-regression.pdf. Accessed: 2024-04-02.

[8] Douglas C Montgomery and George C Runger. *Applied statistics and probability for engineers*. John wiley & sons, 2010.

[9] Nvidia. *Memory Bandwidth*. https://docs.nvidia.com/gameworks/content/technologies/mobile/gles2_perf_mem_bandwidth.htm. Accessed: 2024-05-05.

[10] Ilias Sekkaf. *Computer Parts*. https://www.kaggle.com/datasets/iliassekkaf/computerparts/data.

[11] Statdoe. *Boxplot in R*. https://statdoe.com/one-way-anova-and-box-plot-in-r/. Accessed: 2024-04-20.

[12] Statdoe. *Scatterplot for One Factor in R*. https://statdoe.com/step-by-step-scatterplot-for-one-factor-in-r/. Accessed: 2024-04-20.

[13] testbook. *Multiple Regression: Formula, Theory, and Solved Examples*. https://testbook.com/maths/multiple-regression. Accessed: 2024-05-05.

[14] Hadley Wickham. *tidyverse*. https://tidyverse.tidyverse.org/. Accessed: 2024-04-22.