A logo of a planet

Description automatically generated

**Group Assignment**

**TECHNOLOGY PARK MALAYSIA**

**CT107-3-2-ENTS**

**PROGRAMMING FOR DATA ANALYSIS (PFDA)**

**Web Hacking Trends**

**APD2F2411SE / APU2F2411SE**

**HAND OUT DATE: 22 November 2024**

**HAND IN DATE: 19 February 2025**

**INSTRUCTIONS TO CANDIDATES:**

Complete the assignment and submit it online through Moodle. Your assignment submission is integrated with Turnitin option for plagiarism check.

|  |  |
| --- | --- |
| **NAME** | **TP NO.** |
| Teoh Wei Lun | TP079355 |
| Lee Xin Kher | TP079243 |
| Ng Jiao Zhi | TP065446 |
| Kareshma Kaur A/P Pardeep Singh | TP071385 |

Table of Contents

[1.0 Introduction 3](#_Toc190900722)

[1.1 Data Description 3](#_Toc190900723)

[1.2 Assumption 4](#_Toc190900724)

[1.3 Hypothesis and Objectives 5](#_Toc190900725)

[2.0 Data Preparation 6](#_Toc190900726)

[2.1 Data Import 6](#_Toc190900727)

[2.2 Data Cleaning 6](#_Toc190900728)

[3.0 Data Analysis 19](#_Toc190900729)

[3.1 To investigate the relationship between the amount of ransom and financial loss caused by hacking. Teoh Wei Lun – TP079355 19](#_Toc190900730)

[3.2 To investigate the relationship between operating system and the amount of financial loss caused by hacking. Lee Xin Kher – TP079243 35](#_Toc190900731)

[3.3 To examine the correlation between downtime and the amount of financial loss resulting from hacking attacks. Ng Jiao Zhi- TP065446 50](#_Toc190900732)

[3.4 To investigate the relationship between the web server and financial loss caused by hacking. Kareshma Kaur – TP071385 64](#_Toc190900733)

[4.0 Conclusion 75](#_Toc190900734)

[4.1 Discussion 75](#_Toc190900735)

[4.2 Recommendations 75](#_Toc190900736)

[4.3 Limitation and future direction 75](#_Toc190900737)

[5.0 Referencing 77](#_Toc190900738)

# 1.0 Introduction

Hacking incidences are increasing that results financial losses and operational interruption for companies. Understanding the reasons that cause these losses is important for improving cybersecurity efforts. This study looked at hacking trends over 15 years to find important factors that influence financial loss. Our investigation uses data pre-processing, statistical modeling and visualization tools to investigate the effect of the variables. The insights will assist in developing cybersecurity recommendations for companies.

## Data Description

|  |  |
| --- | --- |
| **Variables** | **Description** |
| Date | The recorded date when the website defacement occurred. |
| Notify | The individual or group that reported the defacement incident. |
| URL | The web address of the compromised website. |
| IP | The IP address of the affected server. |
| Country | The geographical location of the server hosting the defaced website. |
| OS | The operating system running on the compromised server. |
| Webserver | The version of the web server software used by the affected server. |
| Encoding | The character encoding format used in the defacement message. |
| Lang | The language displayed on the defaced webpage. |
| Ransom | The amount paid in ransom (in thousands). |
| Downtime | The number of days the system remained unavailable due to the attack. |
| Loss | The financial loss incurred as a result of the defacement |

## Assumption

Several assumptions about the dataset and its context can be drawn from data cleaning steps:

* Missing values in the "Ransom" column are imputed as 0, assuming no ransom was demanded instead of a data entry failure.
* Missing values in "IP" and "URL" are replaced with "Unknown”, assuming the data was either missing or hidden to maintain data for analysis.
* Rows with more than six missing values are eliminated, assuming much missing data makes the entry undependable for useful analysis.
* Duplicate rows are eliminated, assume duplicates are accidental errors and not valid multiple reports of the identical incident.
* Country names are standardized. Variations such as "unitedkingd" → "United Kingdom" are formatting inconsistencies, not different entities.
* Web servers are standardized. Groups similar versions under general categories such as "Apache/2.4.41" → "Apache", assuming detailed versions are unneeded.
* Different OS versions such as "Windows 10" and "Windows Server 2016" are grouped under major categories like "Windows."
* Empty values and whitespace in non-"Notify" columns are handled as missing values, assuming spaces or empty strings do not contain valid data.
* Final review to confirm dataset accuracy that assume remaining missing values are minor and unimportant to key findings.

## Hypothesis and Objectives

**Hypothesis:**

The amount of financial loss due to hacking attacks are impacted by the following factors, and they are downtime in days when system is unavailable, Operating System (OS) of server, webserver and the ransom amount paid.

**Objective:**

1. To investigate the relationship between the ransom paid with the amount of financial loss as a result of hacking.

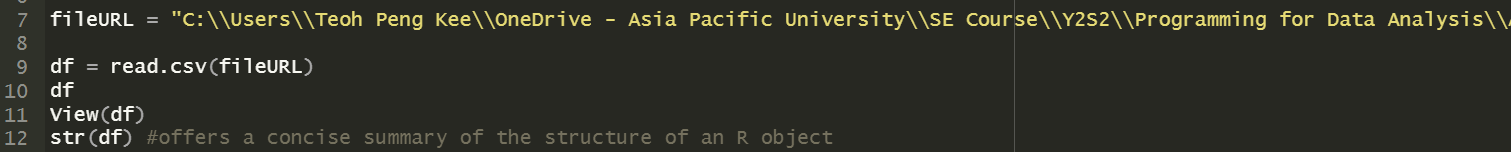
2. To examine the correlation between downtime and the amount of financial loss resulting from hacking attacks.

3. To investigate the relationship between OS and the amount of financial loss caused by hacking.

4. To analyze how the web server impacts the amount of financial loss caused by hacking.

# 2.0 Data Preparation

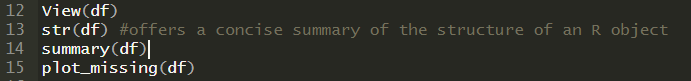
## 2.1 Data Import



The code above imports the dataset that will be applied for this data analysis project. Firstly, change the working directory to the folder containing the dataset excel file titled "hackingData\_OG.csv". Secondly, the data into the R environment using the read.csv function. The dataset has been assigned a variable named data, ‘df’, waiting for pre-processing and data exploration.

## 2.2 Data Cleaning

**Step 1: Checking for Missing Values**

****

Since real-world datasets often contain missing values, it is crucial to identify gaps before proceeding with analysis. A visual representation of missing values is generated to assess which columns have significant data gaps that need attention. This helps in determining the extent of missing data and deciding on appropriate handling strategies.

**Step 2: Replacing Empty Values with Missing Data Indicators**

**A screen shot of a computer screen

AI-generated content may be incorrect.**

Some missing values in the dataset are recorded as empty strings instead of standard missing value indicators. To ensure consistency, all empty values and whitespace-only entries, except in the "Notify" column, are replaced with a uniform missing value format. This standardization makes it easier to manage missing data in subsequent steps.

**Step 3: Rechecking Missing Values**

Once empty values have been replaced, the dataset is rechecked to confirm the changes. Another missing values summary is generated to count the missing entries in each column. This allows for a clear assessment of how much missing data remains and helps in deciding whether to remove certain records or fill missing values with estimated ones.

**Step 4: Removing Rows with Excessive Missing Data**

**A screen shot of a computer

AI-generated content may be incorrect.**

Rows containing a high number of missing values can reduce the reliability of the dataset. To maintain data quality, any row with more than six missing values is removed. After this step, another check is performed to ensure a reduction in missing data and to verify that the retained records still contain meaningful information.

**Step 5: Removing Duplicate Rows**

**A computer screen shot of a black screen

AI-generated content may be incorrect.**

Duplicate rows can distort analysis results and introduce bias. The dataset is examined for duplicate records, which are identified and removed. This ensures that each record is unique, preventing redundancy and maintaining data integrity.

**Step 6: Imputing Missing Values for Certain Columns**

**A screen shot of a computer program

AI-generated content may be incorrect.**

Instead of removing all missing values, specific columns are assigned reasonable default values where appropriate. Missing values in the "IP" and "URL" columns are replaced with "Unknown" to indicate that the information is not available. In the "Ransom" column, missing values are replaced with 0, assuming no ransom was demanded if no data was provided. Additionally, the "Ransom" column is converted into a numeric format for consistency in further analysis.

**Step 7: Standardizing Country Names**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A computer screen shot of a computer program

AI-generated content may be incorrect.**

**A screen shot of a computer

AI-generated content may be incorrect.**

Country names in the dataset often appear in inconsistent formats due to variations in capitalization, punctuation, and extra spaces. To ensure uniformity, all country names are converted to a consistent format, with lowercase text, removed punctuation, and stripped spaces. This prevents errors in geographic data interpretation.

**Step 8: Correcting Country Name Variations**

**A screen shot of a computer screen

AI-generated content may be incorrect.**

Some country names appear in multiple forms due to minor variations in spelling or abbreviations. A predefined mapping is applied to correct these inconsistencies, ensuring that all country names follow a standardized naming convention. This prevents duplicate entries for the same country and enhances accuracy in regional analysis.

**Step 9: Cleaning and Categorizing Web Server, Operating System, Language and Encoding Data**

A screen shot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer screen

AI-generated content may be incorrect.

The dataset includes information on web servers used by hacked websites, but the data is recorded in various formats. To make the data more structured, web servers are categorized into bigger groupings such as "Apache," "Microsoft-IIS," "Nginx," and "LiteSpeed." Security-related web servers are placed in a separate "Security-Focused" category, while unknown values are labeled as "Unknown." For OS, OS are divided into major groups such as "Unix/Linux," "Windows," and "Mac." Same technique applied to Language and Encoding Data.

**Step 10: Using AI model techniques to impute certain data**

A screen shot of a computer

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer code

AI-generated content may be incorrect.

Multivariate Imputation by Chained Equations (MICE) and random forest used to impute loss feature data due to its robustness, iteratively estimating missing values by modeling each variable as a function of the others while maintaining relationships between variables.

**Step 11: Finalise data types of data**

Due to string or NA imputations, often times, data types of certain variables might get changed. In order maintain the data type, downtime data is converted back to initial state as standardized numeric.

**Step 12: Final Verification, Data Review, Clean Data Export**

After all data cleaning steps have been completed, a final review is conducted to ensure that no missing values or inconsistencies remain. A summary of the dataset is generated to provide an overview of its completeness and structure. The final cleaned dataset is visually inspected to confirm that it is ready for further analysis and exported.

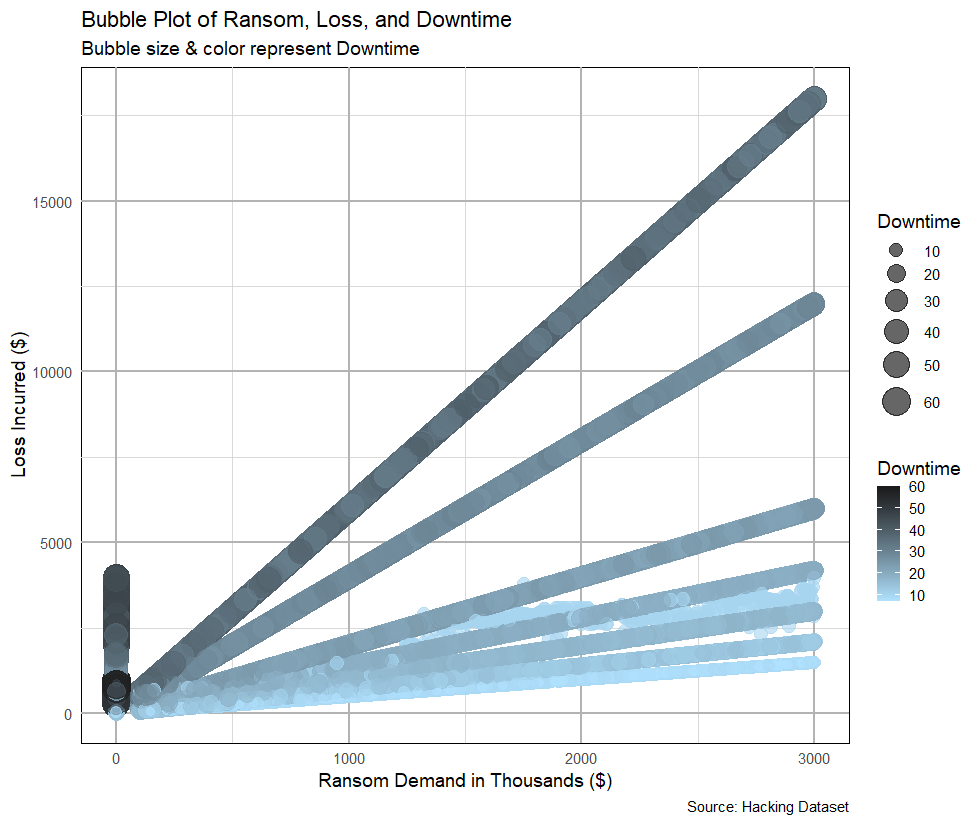
# 3.0 Data Analysis

## 3.1 To investigate the relationship between the amount of ransom and financial loss caused by hacking. Teoh Wei Lun – TP079355

**Question 1: Is ransom the only indicator as independent variable that associated strongly with financial loss?**

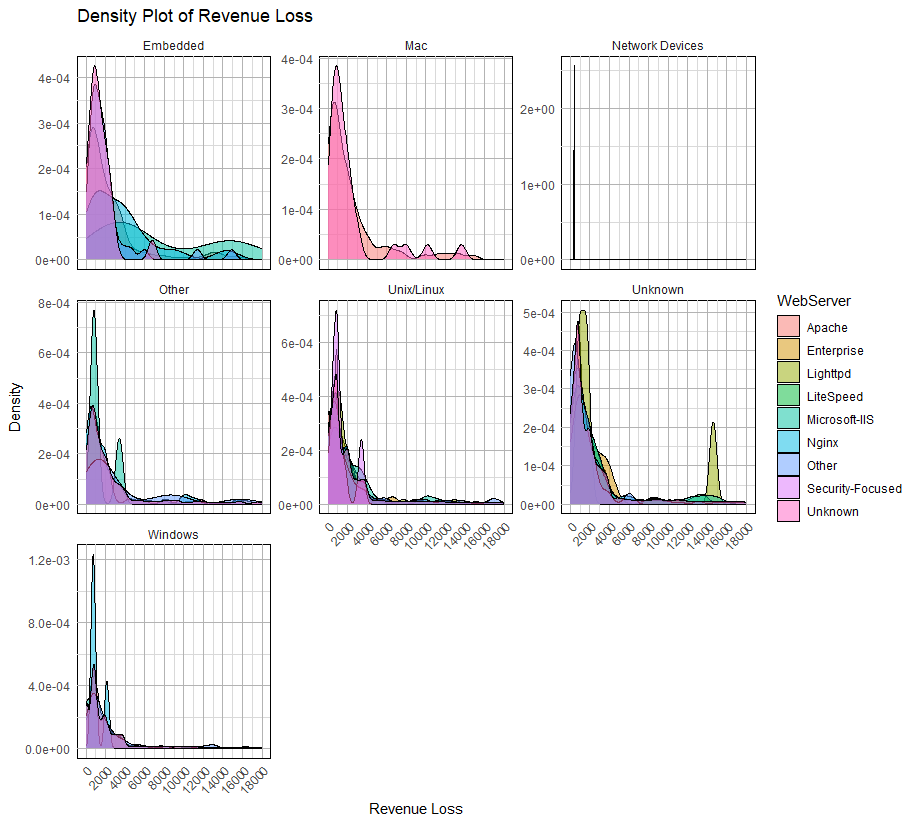
**A screen shot of a computer code

AI-generated content may be incorrect.**



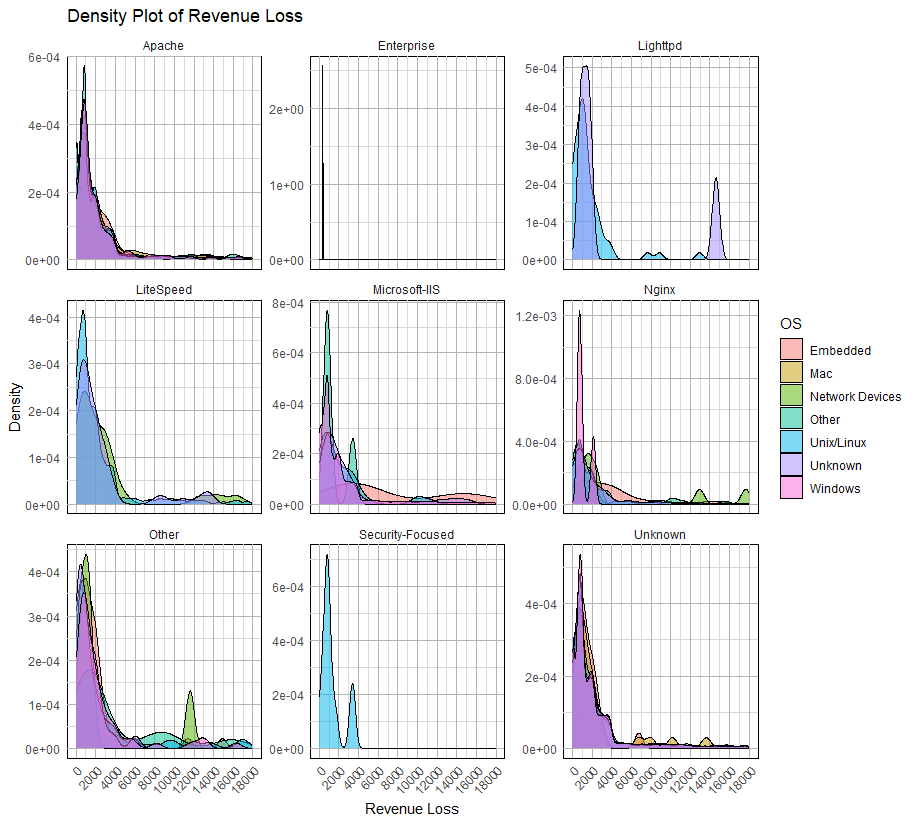
*Figure 3.1.1: Bubble Plot of Ransom vs Loss vs Downtime*

The bubble plot shows the relationship between ransom demand ($) and loss incurred ($), with bubble size and color indicating downtime. The plot shows various linear trends, indicating multiple patterns in ransom-loss correlations (Andy, 2021). The clustering at the origin at indicates the loss incurred less than $5000 and the highest downtime period concentrated here. This may due to the hackers’ interest, trying their capabilities on technicality instead of demanding ransom. The high ransom amount of data are only mostly falls under $3750 loss. In overall, The data demonstrates a substantial positive link between ransom demand and loss, showing that as ransom demands rise, so do financial losses. For example, greater ransom demands frequently result in losses that reach $500,000. Furthermore, study reveals that downtime levels, which are crucial for operational efficiency, rise alongside ransom demand and incurred losses. This relationship highlights the multifaceted impact of ransom demands on financial and operational stability.



*Figure 3.1.2: Density Plot of Revenue Loss on WebServer, faceting OS*

The majority of revenue declines occur at lower values, exhibiting right-skewed distributions among server types. Unix/Linux systems prevail in most areas, whereas Windows systems are mainly found in Microsoft-IIS. Certain servers display bimodal distributions, signifying diverse loss patterns. Severe revenue declines are uncommon but achievable, indicating possible weaknesses. The Security-Focused panel displays unique loss trends, signifying varying security implications

**

*Figure 3.1.3: Density Plot of Revenue Loss on WebServer, faceting OS*

The density plots suggest that Apache and Nginx are the dominant web servers across different device types, with revenue loss figures largely clustered around zero. All distributions show positive skewness, indicating that most losses are on the lower side. Additional peaks in classifications such as Unknown indicate possible anomalies. To effectively reduce revenue loss, strategies must focus on Apache and Nginx, while also delving into categories with secondary peaks to comprehend the reasons behind substantial losses

**Question 2: What is the correlation between ransom & financial loss?**

**A screen shot of a computer

AI-generated content may be incorrect.**

A screenshot of a graph

AI-generated content may be incorrect.

*Figure 3.1.4: Heatmap on Pearson’s Correlation Test for Loss, Ransom and Downtime*

The heatmap indicates a strong positive correlation (0.78) between Ransom and Loss, supporting H₁ (significant association exists). Downtime has a weak positive association (0.15) with Loss but a weak negative correlation (-0.16) with Ransom, indicating a minor impact. The weak relationship between Downtime and Ransom is beneficial where the multicollinearity is minimized. The findings confirm that higher ransom demands result in higher financial losses, but downtime is not a significant influence. Furthermore, the p-values are < 2.2e-16, which is a strong evidence to reject the null hypothesis (H₀).

A computer code on a black background

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

*Figure 3.1.5: Heatmap on Spearman’s Rank for Loss, Ransom and Downtime*

The Spearman correlation heatmap shows a moderate positive correlation (0.53) between Ransom and Loss, which is less than the Pearson correlation, indicating a nonlinear relationship (Mehreen, 2025). Downtime has a larger positive connection (0.38) with Loss, indicating some impact. The slight negative connection (-0.17) between ransom and downtime persists. These findings support H₁ (non-linearity occurs), while rejecting H₀. The data imply that, while greater ransoms result in higher losses, downtime might also contribute, requiring further analysis with non-linear models.

A screen shot of a computer program

AI-generated content may be incorrect.

A graph showing the difference between actual and predictive loss

AI-generated content may be incorrect.

*Figure 3.1.6: Predicted Loss vs Actual Loss on linear regression model on Loss and Ransom*

The Actual vs. Predicted Loss plot shows significant deviations from the ideal (red dashed) line, indicating poor model fit. The diverging blue lines suggest potential issues like heteroscedasticity, multicollinearity, or discrete ransom values, leading to inconsistent predictions. The model struggles to capture variability, implying non-linearity. These violations reject H₀ (linear relationship and homoscedasticity) and support H₁ (model inadequacy and need for transformation or a complex model) for better predictive performance (Dooinn, 2023). Outliers or data structure issues may also contribute.

A graph of loss versus loss

AI-generated content may be incorrect.

*Figure 3.1.7: Residual Plot of Loss with Predictor Ransom*

H₀ (Null Hypothesis): The variance of residuals is constant (homoscedasticity), and a linear relationship exists between ransom and financial loss.

H₁ (Alternative Hypothesis): The variance of residuals is not constant (heteroscedasticity), and the relationship is non-linear.

The funnel-shaped pattern indicates heteroscedasticity, where residual variance increases as fitted values rise, violating regression assumptions. The LOESS trend line suggests a non-linear relationship, meaning a simple linear model may be inadequate (Becaye, 2023). Additionally, extreme residuals indicate outliers or influential points, potentially distorting predictions. Since heteroscedasticity and non-linearity are evident, we reject H₀ (constant variance and linearity) and accept H₁ (non-constant variance and non-linearity), suggesting a need for data transformation or a more complex regression model.

**Question 3: Can the potential financial loss of a hacking incidents be predicted based on amount of ransom incurred?**

**Experiment 1: random forest model (Loss~ Ransom)**

A black screen with white text

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

A graph of a function

AI-generated content may be incorrect.

The random forest model on the Experiment 1 produces considerable prediction errors. MAE (1315.24) indicates significant deviations, while MSE (3,981,574) and RMSE (1995.39) indicate considerable variability, possibly caused by overfitting, underfitting, or noise. Ransom's removal significantly increases error, as evidenced by a %IncMSE of 843.47% and an IncNodePurity of 1.2526 x 10¹². Despite its significance, large error values indicate that Ransom alone is insufficient to forecast Loss, implying the necessity for additional features for model improvement.

**Experiment 2: random forest model (Loss~ Ransom + Downtime)**

A screenshot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

A graph of a function

AI-generated content may be incorrect.

The enhanced random forest model from experiment 2 has lower prediction errors. MAE (182.53) indicates smaller deviations, while MSE (78,750.42) and RMSE (280.63) indicate lower variability, resulting in improved accuracy. Feature importance study indicates %IncMSE (78.10%) for Ransom and 163.90% for DownTime, implying that minimising DownTime has a greater impact on error. IncNodePurity data reveal that both features greatly minimize variance, with Ransom having a greater impact. Smaller errors imply better model performance, although further tuning or feature additions may improve predictions even further.

**Experiment 3: random forest model (Loss~ Ransom + Downtime)**

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

This random forest model from experiment 3 has worse prediction accuracy than the prior model. MAE (311.01) has increased from 182.53, showing greater average errors. MSE (187,110.8) and RMSE (432.56) have both increased from 78,750.42 and 280.63, indicating larger volatility and poorer generalization. Feature importance highlights: %IncMSE (51.37%) for Ransom, 62.50% for DownTime, and 1.98% for WebServer, with DownTime having the highest influence. IncNodePurity validates Ransom as a significant predictor, followed by DownTime and lastly is WebServer.

**Comparison between random forest models from experiments conducted**

A computer screen shot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

RF (Loss ~ Ransom) has the largest error, with MAE of 1315.24 and RMSE of 1995.39, indicating poor predictions. Adding DownTime in RF (Loss ~ Ransom + DownTime) improves performance, lowering MAE to 182.53 and RMSE to 280.63. Including WebServer in RF (Loss ~ Ransom + DownTime + WebServer) reduces accuracy, raising MAE to 311.01 and RMSE to 432.56. This suggests that Downtime is an essential predictor, whereas WebServer may create noise or unnecessary complication.

**Question 4: What strategies can organizations implement to minimize financial losses and operational downtime caused by ransomware attacks, based on the ransom-demand and loss correlation trends observed?**

According to the data analysis, the strong positive correlation (0.78) between ransom requests and financial loss indicates that greater ransom demands usually result in substantial financial repercussions, whereas downtime, despite its weaker link to ransom, is vital in revenue loss. Studies on feature importance show that minimizing downtime greatly enhances model accuracy, emphasizing its effect on financial stability. In light of these results, organizations ought to implement a multi-tiered strategy to successfully reduce losses associated with ransomware.

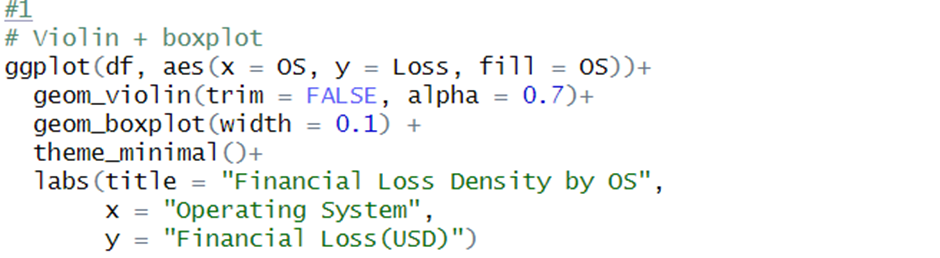
To improve cybersecurity, businesses ought to actively bolster their security systems by implementing advanced endpoint detection and response (EDR) tools, frequently updating software, and addressing vulnerabilities to avert unauthorized access, while also utilizing network segmentation to restrict the proliferation of ransomware attacks. Moreover, creating an all-encompassing backup plan is crucial, which entails keeping regular, encrypted, and unchangeable backups to enable swift recovery without the need for ransom, while also housing backups in isolated settings distinct from the primary network to avert corruption.

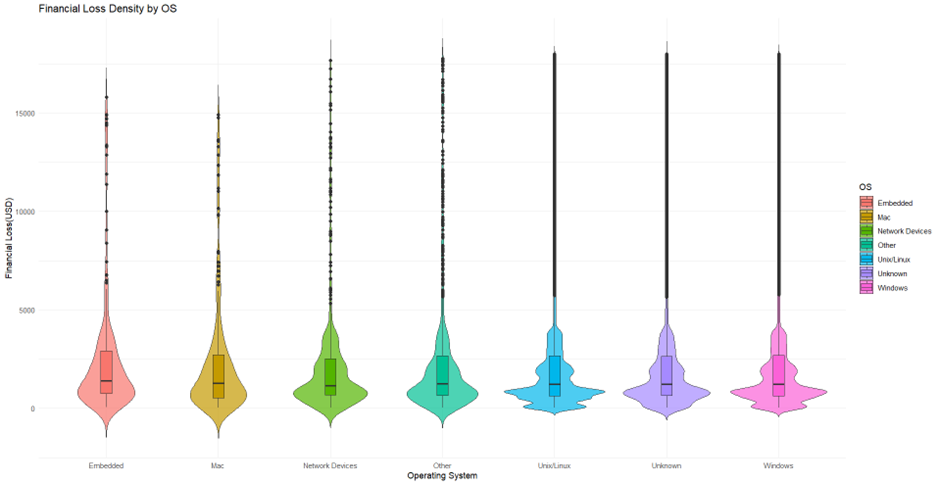
Furthermore, organizations ought to enhance incident response and minimize downtime by formulating a clear incident response plan that includes ransomware containment and recovery methods, performing routine cybersecurity exercises to guarantee that employees and IT teams can respond proficiently, and funding automated disaster recovery tools to quickly restore systems and lessen operational interruptions. Utilizing predictive and AI-powered security frameworks can enhance security protocols, as machine learning algorithms are capable of spotting irregularities and forecasting possible assaults prior to their intensification, while AI-driven threat intelligence can pinpoint high-risk attack avenues.

Moreover, implementing employee awareness and training initiatives is crucial because ongoing cybersecurity education can diminish the likelihood of phishing and various social engineering attacks, while creating transparent reporting systems for suspicious behaviors can improve early threat identification. By adopting these strategies, organizations can not only minimize financial losses but also lessen operational downtime, thus enhancing their overall resilience against ransomware attacks.

## 3.2 To investigate the relationship between operating system and the amount of financial loss caused by hacking. Lee Xin Kher – TP079243

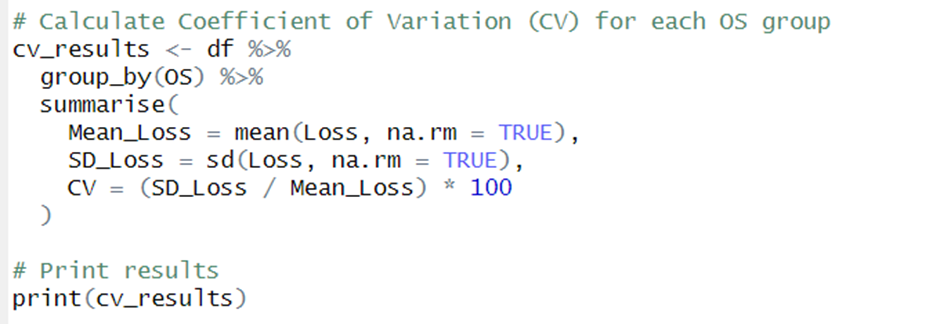
**Question 1: How does the financial impact of hacking incidents vary across different operating systems?**

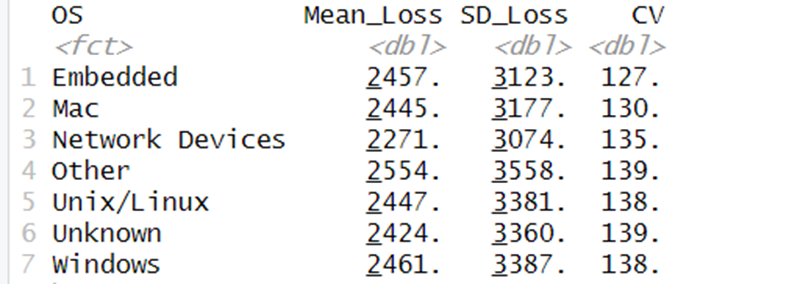




*Figure 3.2.1:Violin plot+Boxplot code and result*

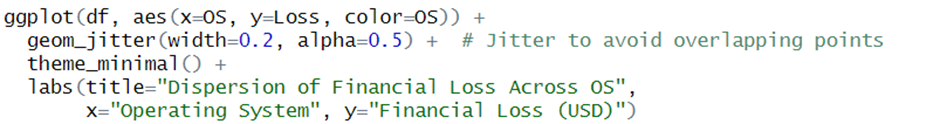
The violin plot with boxplots depicts how financial losses from hacking incidents differs among operating system(OS). The distribution curves show the Windows, Unix/Linux, “Other” and Unknown OS have the greatest range of financial losses, indicating that the expense of hacking incidents on these systems can very greatly. These systems also feature many high-value outliers, showing that some attacks on them cause severe financial damage. This means that they are targeted more frequently or have more weaknesses that result in greater financial harm when hacked. Network devices, Mac and Embedded systems have more compacted distributions that showed that financial losses are more predictable and steadier in general.

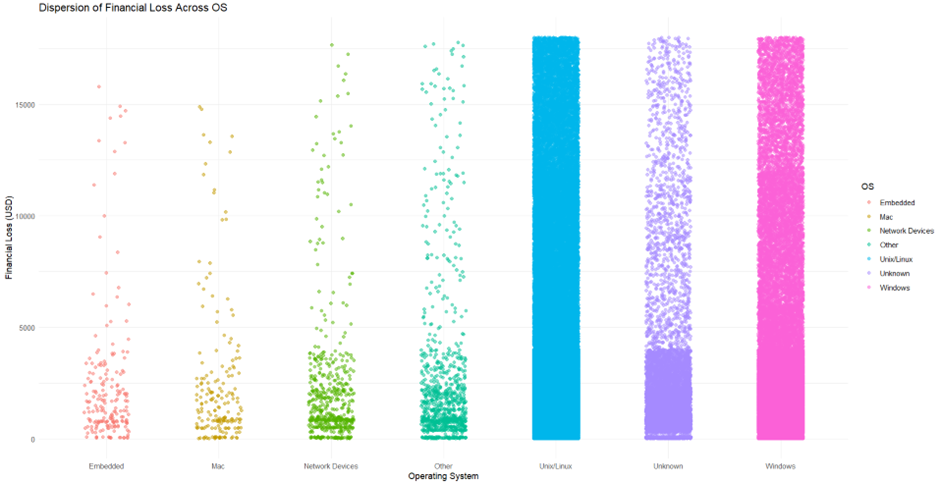




*Figure 3.2.2:CV code and result*

The Coefficient of Variation(CV) is a measure of how much financial losses differ in comparison to the average loss for each OS . The results show “Other” and “Unknown” OS types have the greatest CV(139%), followed by Unix/Linux(138%) and Windows(138%). This shows that hacking incidents are affecting these OS that resulting a wide range of financial losses from slight disruption to significant financial loss. Network devices also show high variability(135%) showing unpredictable financial risks. Embedded systems(127%) and Mac(130%) have slightly lower CV values. It shows that financial losses for these systems are more consistent with but still vary significantly. However, no OS has a low CV meaning all systems suffer large financial losses.



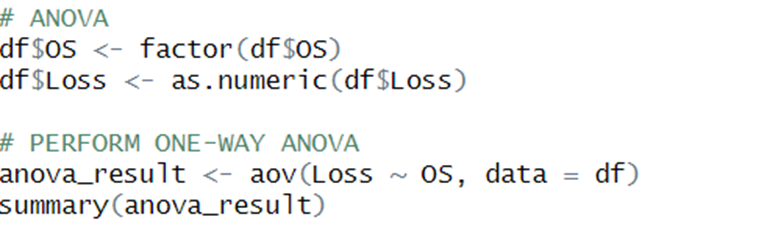


*Figure 3.2.3:Scatter plot and code*

The scatter plot illustrates how financial damages from hacking vary across OS. The graphic shows that all OS suffer both minor and major financial losses, but some have more severe cases. Unknown, Unix/Linux, Windows and “Other” systems have the biggest spread. This is consistent with the coefficient of variation(CV) data, which showed high variability in losses for several OS types.

**Question 2: Do certain operating systems experience significantly higher financial losses compared to others?**

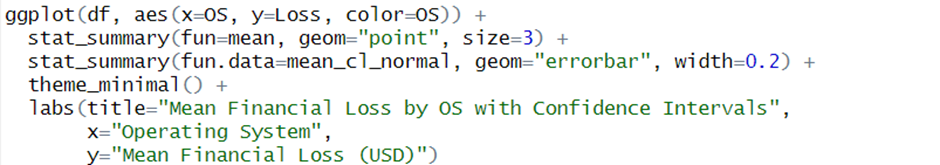
For this question, a one-way ANOVA test is conducted to determine whether certain OSs experience significant higher financial losses compared to others. ANOVA(Analysis of Variance) is a statistical approach for comparing the mean of various groups and determining whether differences are statistically significant (Mackenzie, 2024).

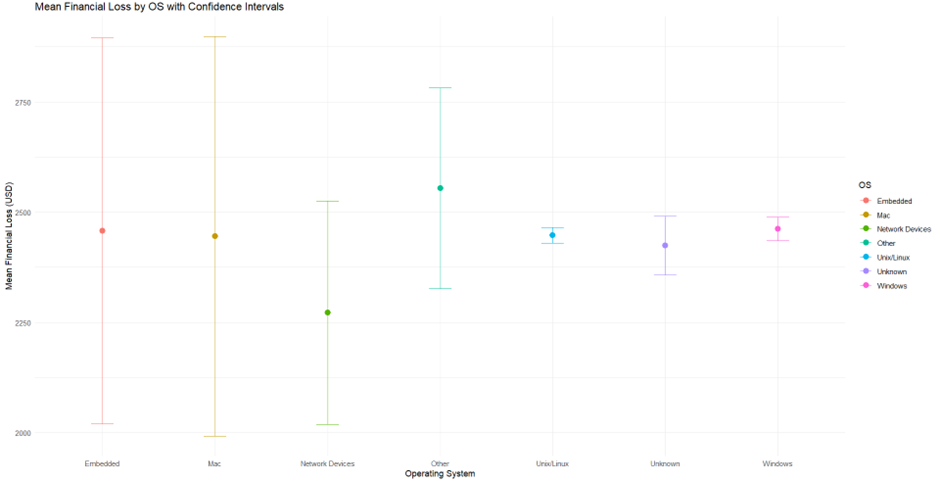




*Figure 3.2.4: ANOVA code and result*

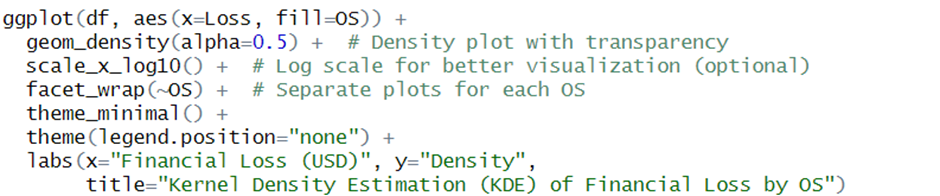
The findings of the one-way ANOVA test show that the OS has no significant effect on financial losses from hacking incidents. The p-value of 0.695 is significantly higher than common standard of 0.05. It shows that any detected differences in financial loss between OS types are more likely related to random variation than an actual influence. Furthermore, the F-value of 0.644 confirms the variation within each OS is much greater than the difference across groups. This implies that the type of OS chosen has no influence on financial loss.

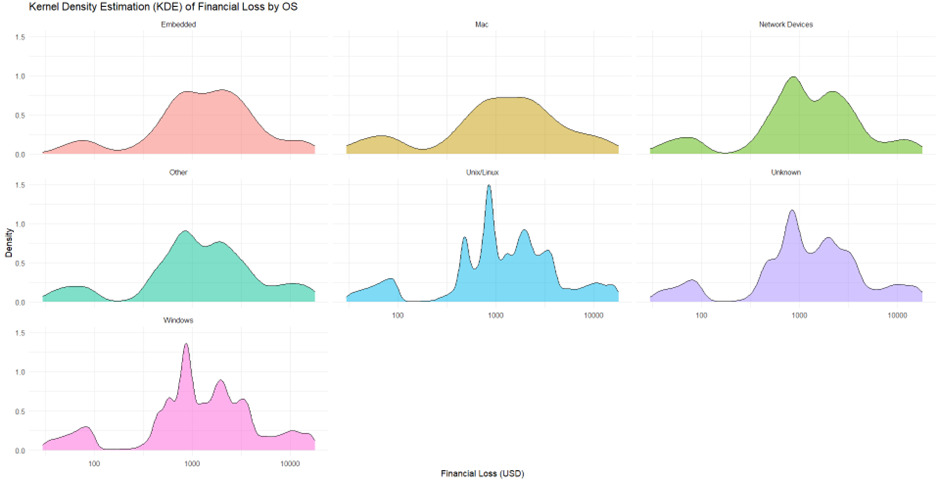




*Figure 3.2.5:Confidence Intervals code and graph*

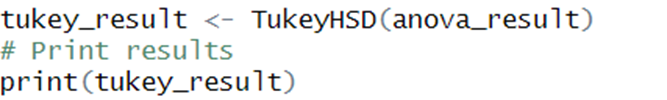
The graph’s confidence intervals provide additional evidence for this conclusion. The intervals for several OS types overlap significantly. This shows that no single OS typically maintains more financial losses than others. While certain OS, such as Unix/Linux and Windows seem to have slightly reduced variability, others such as Mac and Embedded systems display larger confidence intervals, showing greater uncertainty in financial loss estimations. In conclusion, the ANOVA test results shows that there is no significant difference in financial losses across various OS(p-value=0.695).

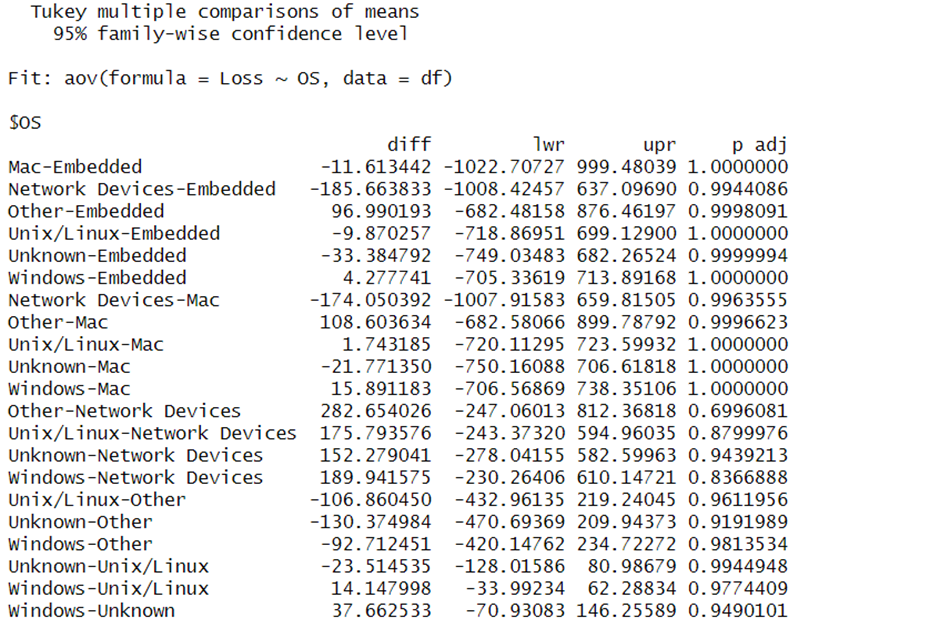




*Figure 3.2.6:Density facet wrap code and graph*

The density graph shows a Kernel Density Estimation(KDE) of financial losses spread into discrete facets with each OS type shown separately. This faceted technique allows for easier understanding comparison of result without overlap found in the first graph. From the results, Windows and Unix/Linux systems show substantial unpredictability and different types of loss values but Mac and embedded systems face more consistent and predictable financial losses. The “Unknown” category seems to combine features from various OS types that result in an inconsistent density distribution.



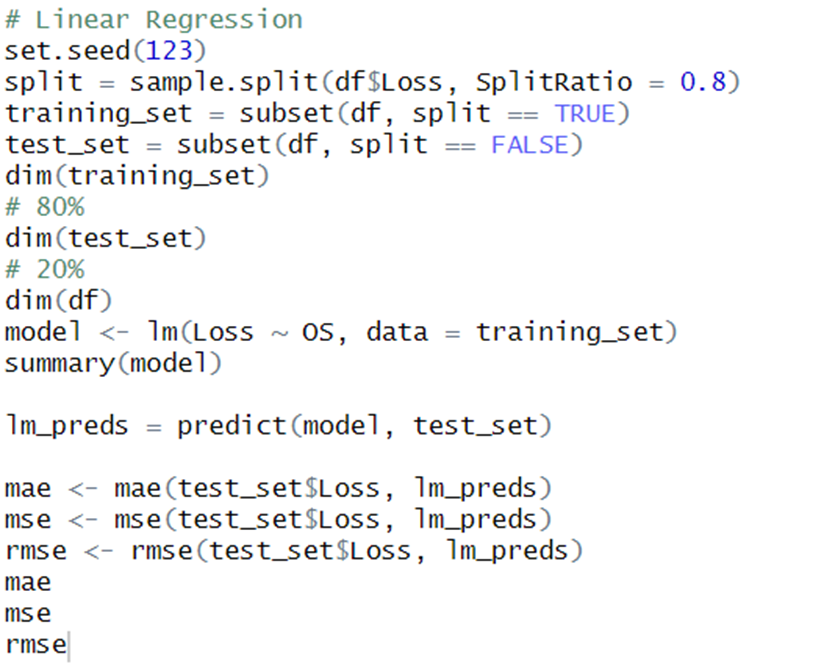


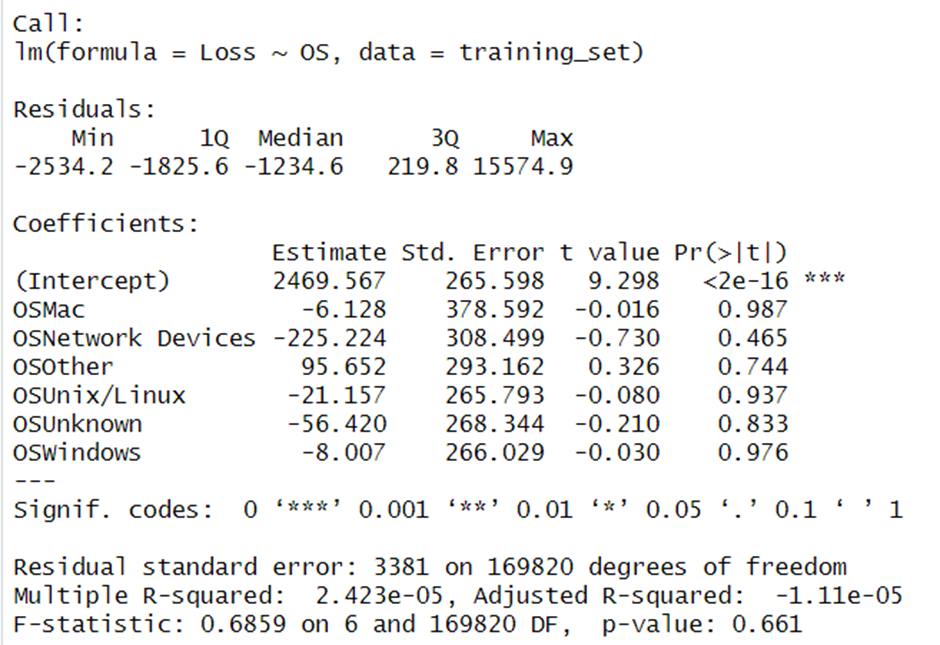
*Figure 3.2.7:Tukey code and result*

The Tukey’s Honest Significant Difference(HSD) test shows that there are no significant differences in financial losses between OS. All pairwise comparisons generated p-values above 0.05. Even the largest detected difference between “Other” and “Network Devices”, is not statistically significant. These findings are consistent with the ANOVA results.

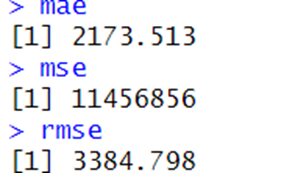
**Question 3: Can organization estimate the potential financial loss of a hacking incident based on different operating systems?**

Linear regression is a regression model that describes variable relationships using a straight line(Kanade, 2023). It finds the line of best fit through your data by looking for the regression coefficient(s) with the lowest total error. It helps in predictions and understanding the importance of this relationship.





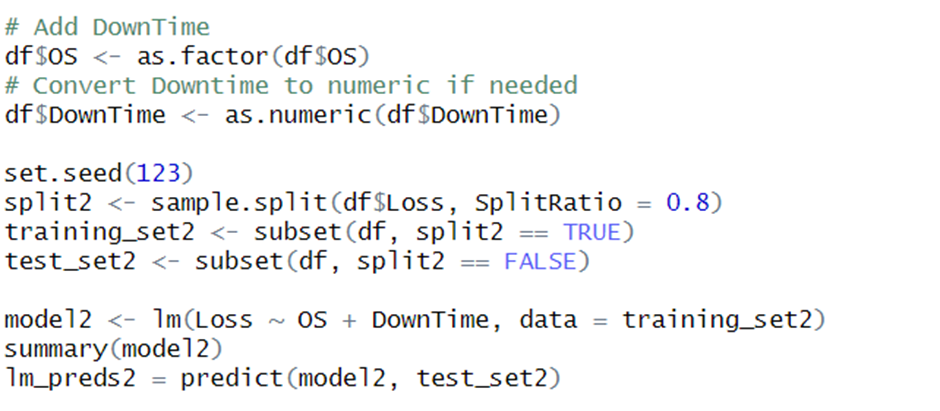
*Figure 3.2.8:Linear regression model1 OS and loss*

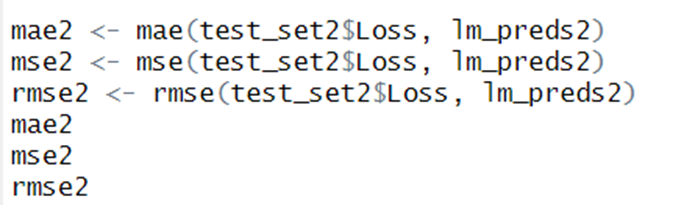


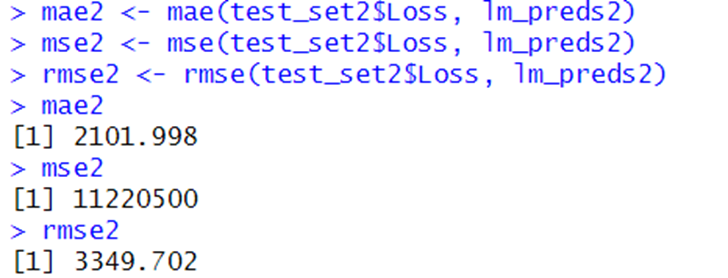
*Figure 3.2.9: MAE,MSE,RMSE result*

The linear regression study determines if the OS can predict financial losses in hacking occurrences. The model’s results show OS is not an important factor. The p-values for all OS categories are substantially above the conventional level of importance(0.05) with the overall model p-value of 0.661. This shows no meaningful connection between OS and financial loss. Furthermore, the R-squared value(0.00002423) is exceedingly low that showed that OS predicts nearly none of the variation in financial loss. The large total standard error(3381) and performance metrics-mean absolute error(MAE) of 11,456,856 and root mean squared error(RMSE) of 3384.798 that confirm the poor predictive accuracy. These findings show that financial loss is driven by other factors other than OS.

To confirm OS is not a reliable predictor of financial loss, another independent variable, downtime is added to conduct the second experiment of linear regression model.

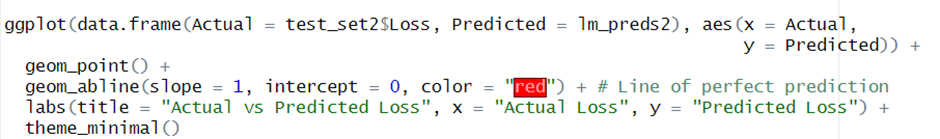








*Figure 3.2.11:Linear regression model2 OS+Downtime and Loss*

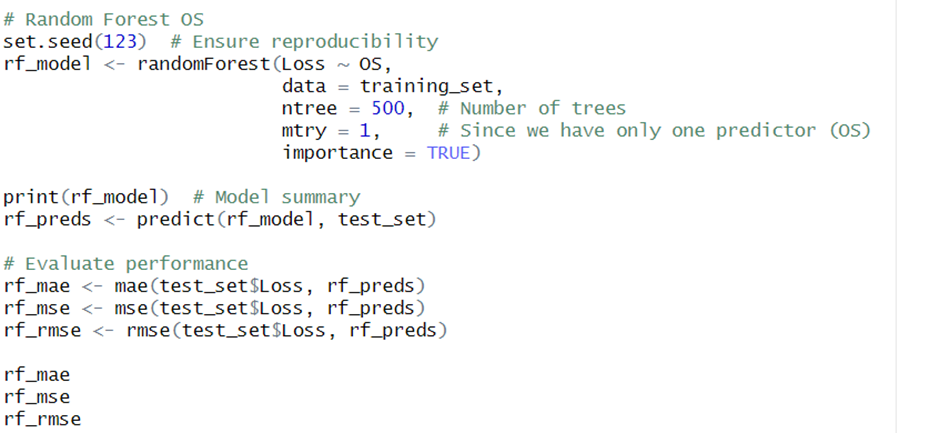


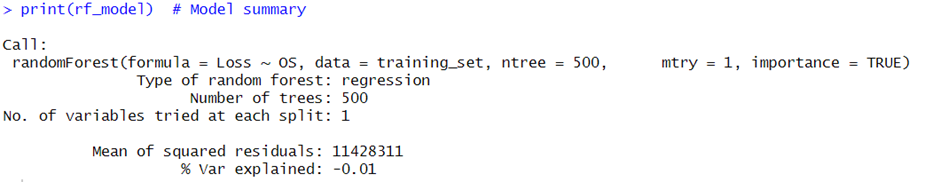


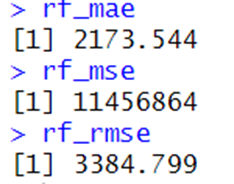
*Figure 3.2.12:Scatter graph with reference line model2*

This linear regression model predicts financial loss depending on the OS and downtime. The analysis graph shows a important connection between downtime and loss. For each unit of downtime, the model predicts an increase in loss by approximately 34 units. These findings show downtime is large influence on loss. However, the model showed no significant link between OS and loss. The model found no statistically significant difference in corresponding expenses between each type of OS. Finally, the model’s overall prediction strength is low.

Therefore, Random Forest is used to compare prediction accuracy with linear regression model. Random Forest is a popular supervised machine learning technique that may be applied to both classification and regression tasks (M, 2024).

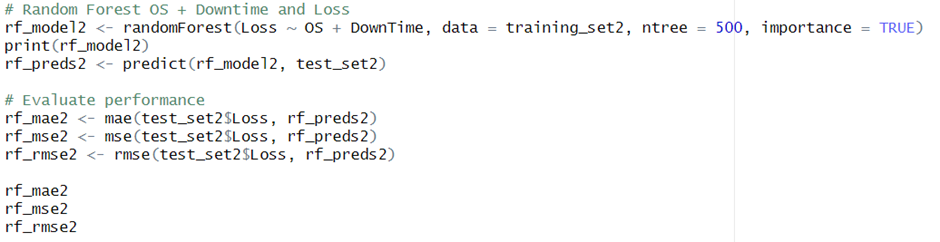


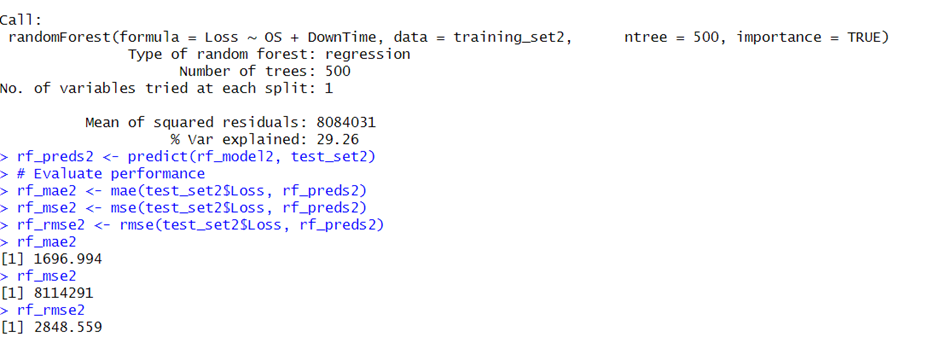




*Figure 3.2.13: Random Forest code OS and Loss*

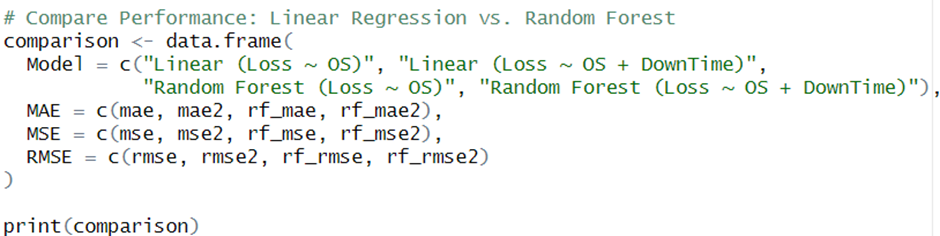
The data shows that the OS used has minimal to no effect on financial loss. Linear regression and Random Forest models provided comparable performance indicators that includes MAE(~2173), MSE(~11,456,864) and RMSE(~3384). This demonstrates that predicting financial loss only based on OS produces inaccurate outcomes. The findings are like the linear regression model results. These findings indicate that OS is not a significant factor influencing financial loss.

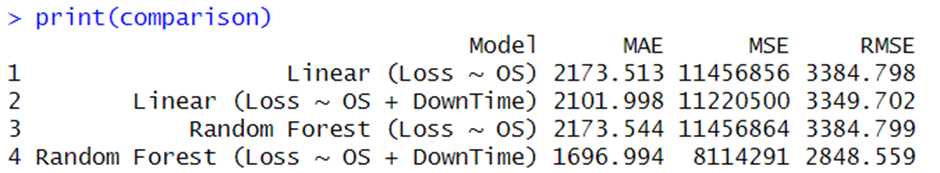




*Figure 3.2.14:Random Forest code OS+Downtime and Loss*

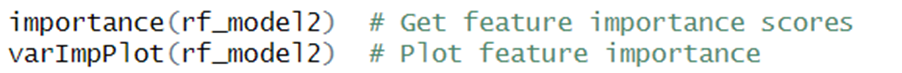
By including both OS and downtime into the Random Forest model, there is a considerable increase in prediction accuracy. The MAE fell from 2173 to 1696, while RMSE decreased from 3384 to 2848. This shows that the model’s projections are getting closer to the real financial loss statistics. Additionally, the variance explained increased from -0.01% to 29.26%, suggesting the model is now responsible for over 30% of the variation in financial loss. This shows that downtime is an important factor affecting financial loss, but OS alone had little to no impact.

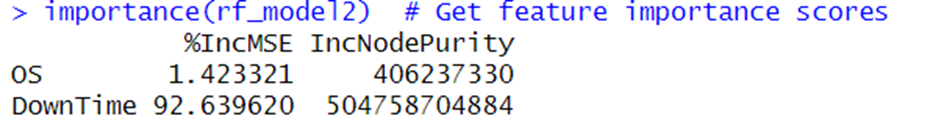


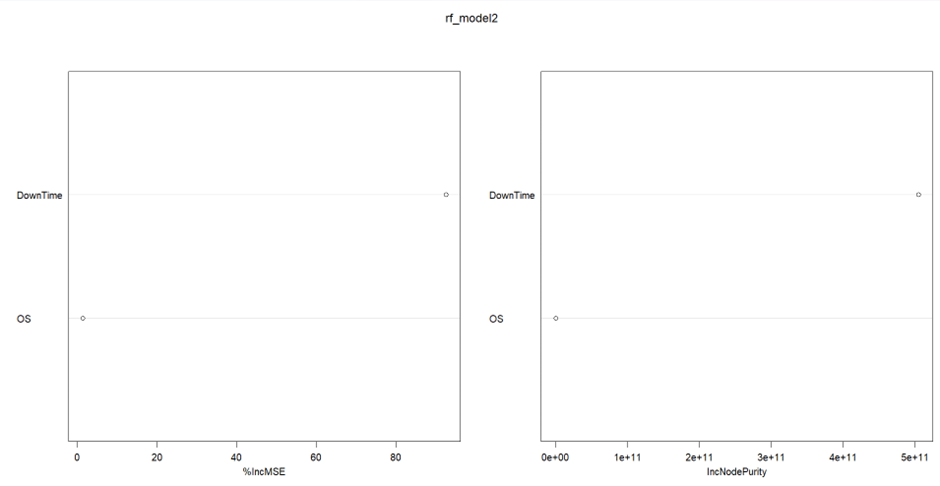


*Figure 3.2.15: Comparison Random Forest and Linear Regression model*

The Random Forest model with OS and downtime performed the best that decreases mistakes and explaining for 29.26% of variance in loss. This means that downtime causes significant financial loss, whereas OS does not. Random forest is more effective than linear regression at capturing complicated patterns.







*Figure 3.2.16:Importance graph*

The feature importance results show downtime has a bigger effect on loss than OS. The%IncMSE(percentage increase in Mean Squared Error) metric shows how the model’s error rate rise if the feature was eliminated. Downtime has a 92.65%IncMSE that shows it is extremely important for predicting loss, but OS has only 1.42% means that it has minimal effect. Similarly, the IncNodePurity(a measure of how much each feature contributes to decrease impurity in decision trees) is significantly higher for downtime(504.76billion) than OS(406billion). This demonstrates that downtime is the more important factor affecting financial loss compared to OS.

**Question 4: What preventive measures and response strategies can organizations implement to minimize financial damage from hacking attempts on operating system(OS)?**

Even though no specific OS was discovered to cause significant higher financial losses, enterprises must secure all OS. Firstly, companies should focus on reducing system downtime. The second linear regression model found that downtime is a strong predictor of financial loss compared to OS. Organizations should upgrade their OS on a regular basis to include security fixes. Hackers frequently target outdated systems, thus keeping OS versions up to date helps to close security gaps and decreases the chance of attack. Strengthening the OS by disabling unnecessary services and shutting open ports makes it more difficult for attackers to break in. Moreover, secure the OS from illegal access by implementing multi-layered security mechanisms such as firewalls, intrusion detection systems(IDS) and encryption to block illegal entry. Strict user access controls and endpoint security tools, such as antivirus software, help to further protect the OS.

## 3.3 To examine the correlation between downtime and the amount of financial loss resulting from hacking attacks. Ng Jiao Zhi- TP065446

**Question 1: What is the overall distribution of downtime and loss in the hacking dataset?**

A screenshot of a computer code

AI-generated content may be incorrect.

*Figure 3.3.1: Visualize data using Hexbin plot.*

A graph of a downtime and loss

AI-generated content may be incorrect.

*Figure 3.3.2: Hexbin plot.*

Technique: geom\_hex()

Analysis Type: Descriptive Analysis

From the graph, the bright yellow region is concentrated at minimal loss values and low downtime around 0-10 days. This suggests that most systems experience short downtimes with low financial losses. A large cluster of points exists at 30-40 days downtime with a wide range of loss values which are $5000 to $15,000 above. This suggests that many systems experienced downtimes around 30–40 days, potentially due to a common maintenance cycle or failure event. Also, very few data points exceed 50 days of downtime and most of these have low to moderate financial loss as clusters in green. The hexbin plot enhances result interpretation by addressing scatter plot clutter in large datasets (Iqra, 2023).

**Question 2: Is the relationship between downtime and loss statistically significant?**

A screenshot of a computer program

AI-generated content may be incorrect.

*Figure 3.3.3: The code of analysis of the impact of downtime on loss.*

A screenshot of a computer code

AI-generated content may be incorrect.

A graph of a loss by a low and high

AI-generated content may be incorrect.

*Figure 3.3.4: The output analysis of the impact of downtime on loss.*

Techniques: cor.test(), wilcox.test(), cliff.delta(), quantile()

Analysis Type: Statistical Analysis, Correlation Analysis, Non-Parametric Statistical Analysis, Group Comparison, Effect Size Analysis, Group Difference Measurement, Descriptive Statistics, Distribution Analysis

The Pearson correlation coefficient (r = 0.146) demonstrates a modest but statistically significant positive association between downtime and loss, implying that as downtime increases, so does loss, although only slightly.

The Mann-Whitney U test, with an exceptionally small p-value (<2.2e-16), indicates that the distribution of loss differs significantly between the high and low downtime groups, demonstrating that experiencing longer downtime tends to result in significantly greater losses. This non-parametric test supports the claim that the discrepancy is not attributable to chance (Intellectus, 2025). The boxplot demonstrates the higher median loss and wider dispersion in the high downtime group, indicating that more downtime leads to greater financial losses.

Cliff’s Delta, with a value of 0.428, suggests a medium effect size, meaning that there is a substantial practical difference in loss between the two groups, those with high downtime tend to have significantly higher loss compared to those with low downtime.

Quantile analysis highlights that the median loss is substantially higher for the high-downtime group (2062) compared to the low-downtime group (872), emphasizing the real-world impact of increased downtime.

Among the four techniques, the Mann-Whitney U test and Cliff’s Delta together provide the most comprehensive analysis of the relationship between downtime and loss, as they not only confirm statistical significance but also reveal differences in distribution and effect size, while Pearson correlation quantifies the linear association, and quantile analysis offers deeper insight into how loss values vary across different downtime categories.

**Question 3: Does any other factor that would help improve the accuracy of downtime and loss predictions?**

A screenshot of a computer program

AI-generated content may be incorrect.

*Figure 3.3.5: Linear Regression (Loss ~ Downtime).*

A screenshot of a computer

AI-generated content may be incorrect.

*Figure 3.3.6: The output of linear regression (Loss ~ Downtime).*

Techniques: lm(), predict(), summary()

Analysis Type: Predictive Analysis, Regression Analysis

The analysis reveals that downtime has a statistically significant impact on financial loss, with each additional day of downtime increasing the loss by approximately $33.97. However, the low R-squared value (2.14%) indicates that downtime alone does not adequately explain the variation in loss, implying the necessity for additional variables. Despite this constraint, the F-statistic (4627, p-value < 2.2e-16) supports the model's overall relevance. The high residual standard error (3345) indicates significant variability, implying that other factors should be investigated to improve the model's forecast accuracy.

A close-up of a white background

AI-generated content may be incorrect.

*Figure 3.3.7: The regression metrics (Loss ~ Downtime).*

A screenshot of a computer code

AI-generated content may be incorrect.

*Figure 3.3.8: The output of regression metrics (Loss ~ Downtime).*

Techniques: mae(), mse(), rmse()

Analysis Type: Model Evaluation, Performance Metrics

The high MAE (2097.521) and RMSE (3363.133) indicate considerable differences between predicted and actual values, while the high MSE (11,310,662) confirms large errors in some predictions. Combined with the low R-squared (0.02141), these findings suggest that downtime alone is a weak predictor of loss. To enhance accuracy, incorporating additional factors influencing revenue loss is essential for a more reliable predictive model.

A screenshot of a computer code

AI-generated content may be incorrect.

*Figure 3.3.9: Multiple regression metrics (Loss ~ Downtime + Ransom).*

A screenshot of a computer program

AI-generated content may be incorrect.

*Figure 3.3.10: Output of multiple regression metrics (Loss ~ Downtime + Ransom).*

Techniques: lm(), predict(), summary()

Analysis Type: Predictive Analysis, Regression Analysis

The summary of the linear regression model (lm\_model\_multi) indicates that both downtime and ransom are highly significant predictors of loss, with p-values < 2e-16. The model has a high R-squared value of 0.6806, indicating it explains approximately 68.06% of the variance in loss. The MAE of 1230.228 indicates the average absolute difference between predicted and actual loss values. The MSE of 3,633,688 and RMSE of 1906.224 represent the average squared and square root of squared differences, respectively (Zach, 2022).

A screenshot of a computer program

AI-generated content may be incorrect.

*Figure 3.3.11: Random forest model (Loss ~ Downtime).*

A screenshot of a computer code

AI-generated content may be incorrect.

*Figure 3.3.12: Random forest model (Loss ~ Downtime).*

Techniques: randomForest(), predict(), importance(), varImpPlot()

Analysis Type: Predictive Modeling, Machine Learning (Random Forest Regression)

The model includes 500 trees and captures a variety of metrics (Niklas, 2024). The evaluation on test data shows a MAE of 1658.44, MSE of 8,067,496.78, and RMSE of 2840.33. These metrics suggest that the model has moderate predictive accuracy, with some variability in predicting loss based on downtime.

A screenshot of a computer code

AI-generated content may be incorrect.

*Figure 3.3.13: Random forest model (Loss ~ Downtime + Ransom).*

A screenshot of a computer code

AI-generated content may be incorrect.

*Figure 3.3.14: Output of Random forest model (Loss ~ Downtime + Ransom).*

The output presents the regression performance metrics of the Random Forest model, evaluating the prediction accuracy for loss. The MAE of 183.68 suggests that, on average, predictions deviate from actual values by this amount. The MSE of 79,373.38 reflects the squared differences between predicted and actual values, giving more weight to larger errors. The RMSE of 281.73 provides an interpretable measure of error magnitude in the same unit as loss, indicating a moderate level of prediction error (Raghav, 2024). These metrics suggest the model has a fair predictive performance.

A computer screen shot of a computer code

AI-generated content may be incorrect.

*Figure 3.3.15:Comparison facet plot.*



*Figure 3.3.16:Output of comparison facet plot.*

Techniques: data.frame(), facet\_wrap(), ggplot(), geom\_point(), geom\_abline()

Analysis Type: Model Comparison, Performance Evaluation

The plot compares actual versus predicted loss for two Random Forest models, one using both downtime and ransom variables, and another using only downtime. The first model closely follows the red reference line, indicating high accuracy, while the second shows poor predictions, implying that downtime alone is insufficient. The red dashed line (geom\_abline()) serves as a baseline for comparison, highlighting prediction errors. Including ransom significantly improves accuracy, making it a crucial factor in loss prediction.

A screenshot of a computer program

AI-generated content may be incorrect.

*Figure 3.3.17:Compare different models’ performance.*

A screenshot of a computer code

AI-generated content may be incorrect.

*Figure 3.3.18:Output of compare different models’ performance.*

The model comparison reveals that the Random Forest model predicting loss using both downtime and ransom performs the best, with the lowest MAE (183.68), MSE (79,373.38), and RMSE (281.73). This indicates higher accuracy and less variability in predictions compared to other models. The linear regression models, both with and without ransom, show higher errors, suggesting they are less effective in capturing the relationship between predictors and loss. The inclusion of ransom as a predictor significantly improves model performance, highlighting its importance in predicting loss.

A close up of a text

AI-generated content may be incorrect.

*Figure 3.3.19:Feature importance.*

A close-up of a number

AI-generated content may be incorrect.

*Figure 3.3.20:Output of feature importance.*

A higher value indicates greater influence in reducing prediction error. Ransom (1.29e+12) has a significantly higher importance than downtime (5.94e+11), suggesting that ransom amount is a stronger predictor of financial loss than downtime duration.

A graph of a number of numbers

AI-generated content may be incorrect.

*Figure 3.3.21:Output of feature importance plot.*

Ransom has a significantly higher importance, around 1.29 trillion, compared to Downtime at 5.94 billion. This suggests that Ransom has a stronger influence on Loss than Downtime.

**Question 4: What is the course of action that could be taken to reduce the revenue loss caused by downtime?**

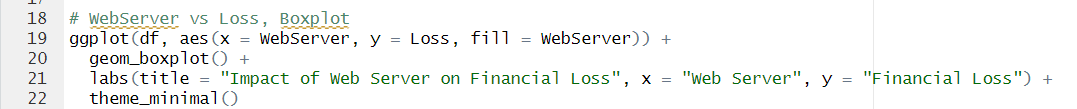
To mitigate revenue loss associated with downtime, prioritize preventive maintenance and rapid incident response to minimize downtime duration. Investing in automated recovery systems, cloud-based redundancies, and enhanced cybersecurity can further reduce downtime-related losses. Additionally, implementing real-time monitoring tools and predictive analytics can help identify potential failures before they occur, enabling proactive intervention (Lauren, 2024).

## 3.4 To investigate the relationship between the web server and financial loss caused by hacking. Kareshma Kaur – TP071385

**Question 1:**

**What is the relationship between Web Server and the financial Loss?**

1st Visualization Technique: Boxplot

****

**A graph of different colored squares

AI-generated content may be incorrect.**

*Figure 3.4.1: Boxplot of Impact of Web Server on Loss*

The boxplot analysis suggests that the type of web server used does not have a significant direct impact on financial loss. The median financial loss and interquartile range (IQR) are relatively similar across different web servers, indicating that no particular server type is consistently associated with higher or lower financial losses. However, there are numerous outliers across all web server categories, especially on the higher end of financial loss, suggesting that organizations using any web server can experience extreme financial losses in certain cases. While some web servers, such as "Unknown" and "Security-Focused," appear to have slightly higher median financial losses compared to others like "Enterprise" and "Lighttpd," the overall distribution remains quite comparable. The wide spread of financial loss values further indicates that factors beyond the choice of web server, such as security practices, cyber threats, and organizational security measures, likely play a more significant role. Therefore, while web server type may have some influence, financial losses are more likely driven by broader security and operational factors.

**2nd Visualization Technique: Bar Plot**

****

**A graph of financial loss

AI-generated content may be incorrect.**

*Figure 3.4.2: Bar plot of Average Loss vs Web Server*

The second visualization technique chosen, the bar chart, illustrates the average financial loss associated with different web servers, providing additional insights into the relationship between web server choice and financial loss. The data suggests that financial loss varies across different web servers, though the differences are not highly dramatic in this technique either. Web servers such as "Apache," "Enterprise," "Microsoft-IIS," "Nginx," and "Other" show relatively high average financial losses, with "Other" having the highest loss. Conversely, "Lighttpd" and "Security-Focused" servers exhibit lower average financial losses, suggesting that these servers may be associated with fewer or less severe incidents leading to financial loss. The "Unknown" category also shows a relatively high average financial loss. While these variations suggest some influence of web server choice, they do not establish a clear causal relationship.

**Question 2:**

**Do the different types of Web Servers significantly impact financial loss?**

Extra feature 1: ANOVA technique

A group of black and white text

AI-generated content may be incorrect.

A close-up of a number

AI-generated content may be incorrect.

*Figure 3.4.3: Results of ANOVA technique*

To determine whether different Web Servers contribute to variations in financial loss, an ANOVA test was conducted. The results showed an F-value of 1.591 and a p-value of 0.122, which is greater than the common significance threshold of 0.05. This indicates that there is no statistically significant evidence to suggest that financial loss varies significantly between Web Servers. In other words, while losses may differ across servers, the variation is likely due to random chance rather than a meaningful relationship between Web Server type and financial loss.

**Question 3:**

**Which specific web servers contribute the most to financial loss?**

Extra feature 2: Kruskal-Wallis test

Extra feature 3: Dunn’s Test with Benjamini-Hochberg correction

A white background with black text

AI-generated content may be incorrect.

A white background with blue text

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

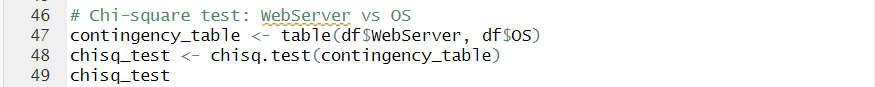
*Figure 3.4.4: Kruskal-Wallis test for Web Server vs Loss*

Since the ANOVA test did not show significant differences, a Kruskal-Wallis test was used as a non-parametric alternative to determine whether the distribution of financial loss varies across Web Servers. The test produced a chi-squared value of 37.041 with a p-value of 1.131e-05, indicating a statistically significant difference in loss distribution. This means that although ANOVA did not find a difference in means, the rank-based Kruskal-Wallis test suggests that some Web Servers experience consistently higher or lower financial losses compared to others. To further explore which web servers contribute the most to financial loss, Dunn's post-hoc analysis with Benjamin-Hochberg correction was performed. The results showed that Apache was associated with the highest financial losses, particularly when compared to LiteSpeed and Microsoft-IIS. Specifically, the difference between Apache and LiteSpeed had a Z-value of 3.25 and a p-value of 0.0138, while the comparison between Apache and Microsoft-IIS had a Z-value of 3.05 and a p-value of 0.0205, both indicating significant financial loss associated with Apache. On the other hand, LiteSpeed and Microsoft-IIS showed lower financial losses in comparison to Apache, with LiteSpeed having a statistically significant reduction in losses when compared to Microsoft-IIS (Z-value of -2.09, p-value of 0.0369). These findings suggest that web servers like LiteSpeed and Microsoft-IIS are less likely to contribute to financial loss compared to Apache.

**Question 4:**

**Is there a significant association between the Web Server and Operating System (OS) used on Loss?**

Extra feature 4: Chi-square test



A screen shot of a computer code

AI-generated content may be incorrect.

*Figure 3.4.5: Chi-square test*

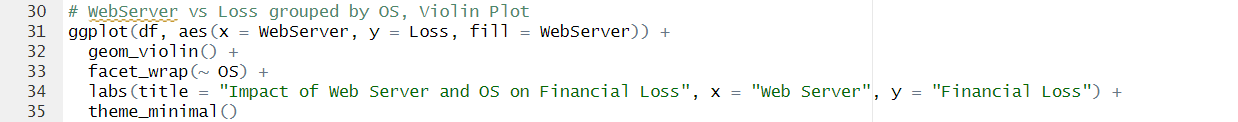
A Chi-Square test was performed to explore whether there is a relationship between Web Server type and Operating System. The test produced an extremely small p-value (< 2.2e-16), which is well below the 0.05 threshold, indicating a strong association between the two variables. This suggests that certain Web Servers tend to be used with specific Operating Systems more frequently than others. However, a warning message in the results indicates that some expected values in the contingency table were too small, which may impact the accuracy of the test.

**Question 5:**

**How does the type of Web Server and Operating System (OS) affect financial Loss?**

**Visualization technique: Violin plot**

Extra feature 5: Violin plot

****

**A graph of different colored objects

AI-generated content may be incorrect.**

*Figure 3.4.6: Violin plot of Impact of Web Server and OS on Loss*

The violin plot analysis indicates that both web server type and OS influence financial loss, with OS having a stronger impact. Financial loss varies across OS categories—Unix/Linux and Windows show wider distributions, while Mac and Network Devices are more concentrated. This suggests Unix/Linux and Windows users may experience both lower and higher losses. Within each OS, financial losses across web servers are similar, though Microsoft-IIS and Nginx show slightly constrained distributions in specific environments like Embedded and Other. Outliers exist across all OS categories, with extreme losses possible regardless of the web server. The "Unknown" OS category shows particularly high losses, implying misclassified systems may pose higher risks. While web server choice matters, OS introduces additional risk factors with greater influence.

**Question 6:**

How is the predictive power of web server to financial loss? How is the contribution of other variables to the prediction?

Extra feature 6: Linear regression model

**1st Experiment: Linear Regression Model for Web Server vs Loss**

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

*Figure 3.4.7: Linear Regression Model*

Based on the results of the predictive analysis using linear regression models, the predictive power of the web server as a factor influencing financial loss appears to be limited. In the first model, which used WebServer as the sole predictor for Loss, the R-squared value was extremely low (0.00004739), indicating that WebServer alone explains very little of the variation in financial loss. Moreover, the F-statistic for this model was not statistically significant (p-value = 0.4317), which suggests that the model does not fit the data well. Most of the individual coefficients for the web server categories also did not show significant effects on the financial loss, with p-values greater than 0.05 for most of them. Only WebServerLiteSpeed showed a significant effect (p-value = 0.0426), although the magnitude of its influence was modest.

When we introduced DownTime as an additional variable in the second model, the predictive power of the model improved slightly, as evidenced by the increased R-squared value of 0.02123. While still low, this indicates marginal improvement. DownTime had a highly significant effect (p < 2e-16), with a strong positive coefficient (33.77), suggesting longer downtimes lead to higher financial losses.

**2nd Experiment: Random Forest Model for Web Server + Downtime vs Loss**

A computer code on a white background

AI-generated content may be incorrect.

A computer screen shot of a program

AI-generated content may be incorrect.

*Figure 3.4.8: Random Forest Model*

The results from the Random Forest model reveal that the variables, WebServer and DownTime, together explain about 29.81% of the variation in financial loss. This indicates that the model captures some significant patterns, although the amount of variance explained is moderate. The Mean Squared Residuals (8003800) suggest that the model’s predictions are fairly accurate, though there is room for improvement.

Further evaluation using performance metrics shows that the Root Mean Squared Error (RMSE) is 2840.739, and the Mean Absolute Error (MAE) is 1675.325, indicating that while the model provides reasonable predictions, some degree of error remains in the forecast. The R-squared value of 0.3014232 further confirms that the model explains 29.81% of the variability in the data, suggesting a moderate predictive power.

Overall, the Random Forest model offers a stronger understanding of the relationship between WebServer and DownTime with financial loss compared to linear regression, highlighting DownTime as a stronger predictor while WebServer contributes through more complex, non-linear interactions.

# 4.0 Conclusion

## 4.1 Discussion

Thorough data analysis has concluded that the hypothesis that the amount of financial loss due to hacking attacks are impacted by downtime, OS of server, ransom and the web server can only be accepted partially. Based on the analysis through multiple techniques, the objectives of the study have been evaluated, and results have proven that Ransom as a factor is the most significant impact on financial loss even through its non-linearity. Downtime has shown a minor but measurable effect, indicating that prolonged service disruptions contribute to increased losses. The operating system(OS) and the web server variables have been concluded to have little to no significant impact on financial loss due to hacking attacks.

## 4.2 Recommendations

Based on the analysis and statistics, some major recommendations can assist reduce financial losses caused by web hacking incidents. Firstly, improve downtime management is important. This may be performed by developing incident handling procedures, installing failover systems and continuously evaluating downtime to reduce financial loss (Barracuda, 2025). Regular checking for security should also be done to identify weaknesses in operating systems and web servers to decrease hacking threats. To prevent ransomware attacks, companies should implement effective mitigation techniques such as regular data backups, strong encryption, employee hacking prevention training on hacker avoidance and a strict no-pay policy for ransom demand. Moreover, web application security must be strengthened through usage of firewalls, input validation and breach detection techniques to protect against common threats. Furthermore, regularly educate employees on cybersecurity concerns to decrease human mistake and avoid hacking and other social engineering attacks. Lastly, companies should consider purchasing cyber insurance for recovery expenses, ransom payments and legal fees in the event of an attack.

## 4.3 Limitation and future direction

**Limitation**

There are some limitations in the current analysis. Outliers may significantly affect the results of regression models and visualizations, yet there is no analysis or handling of outliers in the provided code(Charu C., 2021). While normality tests such as Shapiro-Wilk and Anderson-Darling are used, the code does not address how to manage deviations from normality in residuals, which might invalidate hypothesis test results and impair overall analysis reliability(Mohd Normani, 2022). Furthermore, linear regression models may be overfitting the data, especially if contain too many components, because overfitting occurs when a model learns the noise in the training data rather than the underlying trend. Moreover, the models do not investigate feature engineering strategies such creating interaction terms or algebraic features, which could potentially improve model performance.

**Future Direction**

Several significant changes can be made to increase model performance and dependability. Feature selection such as stepwise selection, Lasso and Recursive Feature Elimination(RFE) help identify the most important aspects influencing financial loss due to hacking to reduce unnecessary data and make the model more precise and interpretation(Dinesh, 2024). Cross-validation like k-fold, stratifies and leave-one-out cross-validation(LOOCV) are used to ensure robust model evaluation to assist in the detection of overfitting and delivering more reliable performance data(Elite, 2022). Besides that, outlier detection using boxplots or Z-scores before model fitting can improve the accuracy of results that allow for the removal or adjustment of outliers that could skewed analysis. The normality of residuals is important since this can be done through data transformations or by using tree-based models like Random Forest or Gradient Boosting which are less sensitive to normality assumptions. Finally, investigating interaction terms, specifically transformations involving variables like Ransom and Downtime, could uncover deeper relationships and improve model findings(Eryk, 2023).

(6697 words)

# 5.0 Referencing

Andy, M. (2021, October 1). *Ultimate Guide to Bubble Charts*. <https://www.netsuite.com/portal/resource/articles/erp/bubble-charts.shtml>

Becaye, B. (2023, April 1). *The Loess Curve: Visualize Trends in your Scatter Plot*. <https://medium.com/@becaye-balde/visualize-trends-in-your-scatter-plot-a0946d6a4299>

Charu C., A. (2021). Outlier Analysis. *Springer.*pg 96-98. <https://sadbhavnapublications.org/research-enrichment-material/2-Statistical-Books/Outlier-Analysis.pdf>

Dinesh, K. (2024, October 15). *A Complete understanding of LASSO Regression*. <https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/>

Dooinn, Ki. (2023, August 17). *How to plot Predicted vs Actual Graphs and Residual Plots*. <https://dooinnkim.medium.com/how-to-plot-predicted-vs-actual-graphs-and-residual-plots-dc4e5b3f304a>

Elite. (2022, July 6). *Overfitting in Machine Learning: What It Is and How to Prevent It*. <https://elitedatascience.com/overfitting-in-machine-learning>

Eryk, L. (2023, April 26). *A Comprehensive Guide to Interaction Terms in Linear Regression*. <https://developer.nvidia.com/blog/a-comprehensive-guide-to-interaction-terms-in-linear-regression/>

Intellectus. (2025). Correlation (Pearson, Kendall, Spearman). <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/>

Iqra, B. (2023, February 4). *Using Hexbin Plots to visualise relationship between two variables*. <https://medium.com/@iqra.bismi/using-hexbin-plots-to-visualise-relationship-between-two-variables-42e26bce3df1>

Kanade, V. (2023, April 3). *What is Linear Regression?- Spiceworks*. Spiceworks Inc. <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/>

Lauren, B. (2024, December 5). How to Minimize Downtime in IT Operations. <https://www.ninjaone.com/blog/how-to-minimize-downtime/>

Lyashenko, V. (2024, April 12). *Cross-Validation in Machine Learning: How to do it right*. neptune.ai. <https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right>

Mackenzie, R. (2024, January 24). *One-Way vs Two-Way ANOVA: Differences, Assumptions and Hypotheses*. Informatics From Technology Networks. <https://www.technologynetworks.com/informatics/articles/one-way-vs-two-way-anova-definition-differences-assumptions-and-hypotheses-306553>

Mehreen, S. (2025). Calculating Spearman’s Rank Correlation Coefficient in Python with Pandas. <https://stackabuse.com/calculating-spearmans-rank-correlation-coefficient-in-python-with-pandas/>

Mohd Normani, Z. (2022). The Limitation of Widely Used Data Normality Tests in  Clinical Research. *Editorial*. <https://avr.tums.ac.ir/index.php/avr/article/view/1027/442>

M, S. (2024, July 23). *Introduction to Random Forest in R*. Simplilearn.com. <https://www.simplilearn.com/tutorials/data-science-tutorial/random-forest-in-r#:~:text=Random%20forest%20is%20a%20popular,the%20performance%20of%20the%20model>.

Niklas, D. (2024, November 26). Random Forest: A Complete Guide for Machine Learning. <https://builtin.com/data-science/random-forest-algorithm>

Raghav, A. (2024, December 17). Know The Best Evaluation Metrics for Your Regression Model ! <https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/>

Zach, B. (2022, December 13). How to Use createDataPartition() Function in R. <https://www.statology.org/createdatapartition-in-r/>