MACHINE LEARNING ASSIGNMENT

"VAH Sample Question Library"

Machine Learning Assignment
Perform PCA and Visualize Data

For steps (i), (ii), (iii), and (iv), refer to the appended code.
From Step (v), PCA using *prcomp (.)* is attempted in the R code, attached in the notepad.

Step 5.1: Justify Scaling Before PCA

Before performing Principal Component Analysis (PCA), it is necessary to standardize the data. PCA works by reducing the dimensionality of the data while retaining as much of the variability as possible. In datasets with multiple variables, each variable might have different units of measurement. Without scaling the data, variables with larger numerical ranges will dominate the analysis, distorting the results. This bias can occur simply because of differences in the scale of the variables, not necessarily because one variable is more important than another.

By scaling the data, we remove this bias. Standardization is achieved by centering the data (subtracting the mean of each feature) and scaling it (dividing by the standard deviation of each feature). This process ensures that PCA focuses on the structure and relationships within the data, rather than being skewed by the units of measurement and their magnitudes.

Step 5.2: Display and Describe the Individual and Cumulative Proportions of Variance

To understand how much variance each principal component (PC) explains in the dataset, PCA outputs both individual and cumulative proportions of variance. This tells us how much of the total data's variability is captured by each PC and the total variance explained as we add more PCs. In the output table, the proportions are displayed up to three decimal places for precision.

Example output table:

Principal Component	Cumulative Proportion
PC1	0.294 (29.4%)
PC2	0.468 (46.8%)
PC3	0.582 (58.2%)

Step 5.3: Outline How Many Principal Components are Adequate to Explain at Least 50% of the Variability in Your Data

To determine how many principal components (PCs) are necessary to explain at least 50% of the variability, we check the cumulative proportion of variance. From the cumulative variance table:

- **PC1** accounts for 29.4% of the variance.
- **PC2** increases this to 46.8%.

"To be used for Academic Reference Only"

- **PC3** brings the cumulative variance to 58.2%, which exceeds the 50% threshold.

Thus, **three principal components (PC1, PC2, and PC3)** are sufficient to explain at least 50% of the variability in the dataset.

Step 5.4.1: Display Coefficients

The coefficients of the principal components can be obtained using the PCA output, which shows how much each original feature contributes to each principal component. These coefficients (also called loadings) are crucial for interpreting the meaning of the principal components. The output should display coefficients up to three decimal places.

Step 5.4.2: Describe Key Drivers for Each Principal Component

Based on the PCA coefficients (loadings), we can determine the key drivers for each principal component. The features with the highest absolute loadings (positive or negative) are the most influential for that component.

PC1 (Principal Component 1):

Key Drivers:

- **Average.Request.Size.Bytes (0.517)**: This feature has a high positive loading, indicating that it contributes significantly to PC1.
- **Attack.Window.Seconds (0.480)**: Another important contributor with a positive loading.
- **IP.Range.Trust.Score (-0.455)**: This feature has a significant negative loading, showing that a lower trust score is associated with this principal component.
- **Average.Attacker.Payload.Entropy.Bits (0.303)**: This feature also contributes positively to PC1.

PC2 (Principal Component 2):

Key Drivers:

- **Hits (-0.532)**: A high negative loading, suggesting that more hits are associated with a lower value on PC2.
- **Average.ping.to.attacking.IP.ms (-0.488)**: Another important negative contributor, indicating that higher ping times lower PC2 values.
- **Average.ping.variability (-0.437)**: Similarly, this feature negatively impacts PC2.
- **Port (-0.346)**: A less significant but still negative contributor.

PC3 (Principal Component 3):

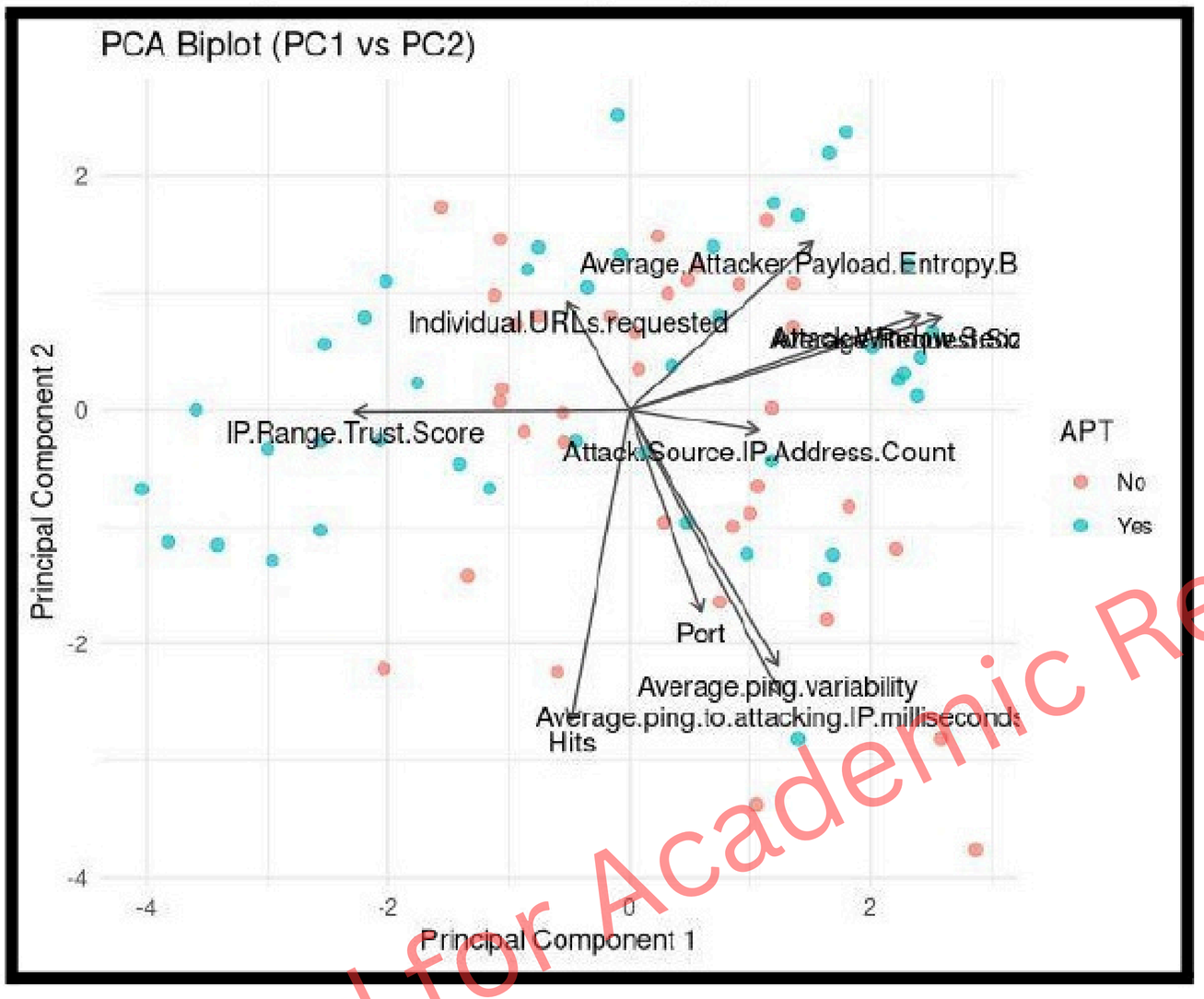
Key Drivers:

"VAH Sample Question Library"

- **Attack.Source.IP.Address.Count (0.446)**: This feature has a positive loading, indicating that more source IP addresses contribute to higher PC3 values.
- **Hits (0.367)**: Positive loading, suggesting that more hits increase PC3 values.
- **Individual.URLs.requested (-0.479)**: A negative loading, indicating that a higher number of individual URLs requested correlates with a lower PC3 value.

It is based on the table in the 5.4 Excel sheet.

(vi): The code is given in the R code file to create and display the biplot for PC1 vs. PC2 to visualize the PCA results.
The code will yield the following image:



"To be used for Academic Reference Only"

Interpretation of PCA Results: PC1 vs. PC2

PCA Plot:

The PCA plot, which visualizes the projection of observations onto the first two principal components (PC1 and PC2), provides insight into how different observations (e.g., APT activity) are distributed in the dataset.

- **Color-Coding**: Observations are color-coded based on their APT status ("Yes" or "No"), allowing us to see whether certain principal components (PC1 and PC2) can distinguish between APT and non-APT activities.
- **Overlap & Distinction**: While there is some noticeable overlap between observations labeled APT = "Yes" and APT = "No," specific regions of the plot are more densely populated by one category. This suggests that PC1 and PC2 can partially separate APT and non-APT activities, but there are still instances where the two categories are similar or difficult to distinguish based solely on these two components.

Loading Plot (Vectors):

The loading plot provides more granular details about which features contribute most to the principal components and how they are related to each other.

- **Magnitude of Influence**: The length of each vector in the loading plot corresponds to the magnitude of influence that a particular feature has on the principal components. Longer vectors represent features that have a stronger influence on PC1 and PC2. For example:
    - **Average.Attacker.Packet.Size** and **Attack.Source.IP.Address.Count** have long vectors, indicating that they significantly contribute to both PC1 and PC2.
- **Correlation Between Features**: The angle between vectors represents the correlation between features.
    - **Small Angles**: When vectors are close together (small angles), it suggests that the features are **positively correlated**. For example, **Average.Attacker.Packet.Size** and **Entropy** might have a small angle between them, indicating they tend to vary in the same direction.
    - **Angles Near 90°**: Vectors that are nearly perpendicular (angles close to 90°) suggest **no significant correlation** between those features. For example, **Attack.Source.IP.Address.Count** and **IP.Range.Trust.Score** might show such an angle, implying that they do not influence each other directly.

Combined Interpretation:

- **APT = "Yes" Observations**: These observations tend to be located along the direction of vectors associated with high values of features like **Average.Attacker.Packet.Size** and **Attack.Source.IP.Address.Count**. These features seem to be more prominent in APT activities, suggesting that larger packet sizes and a higher count of attacking source IP addresses are indicative of APT events.
- **APT = "No" Observations**: Conversely, observations labeled APT = "No" appear in the opposite direction of features such as **IP.Range.Trust.Score**. Higher trust scores are associated with non-APT behaviors, implying that when trust in an IP range is high, the likelihood of an APT event is lower.

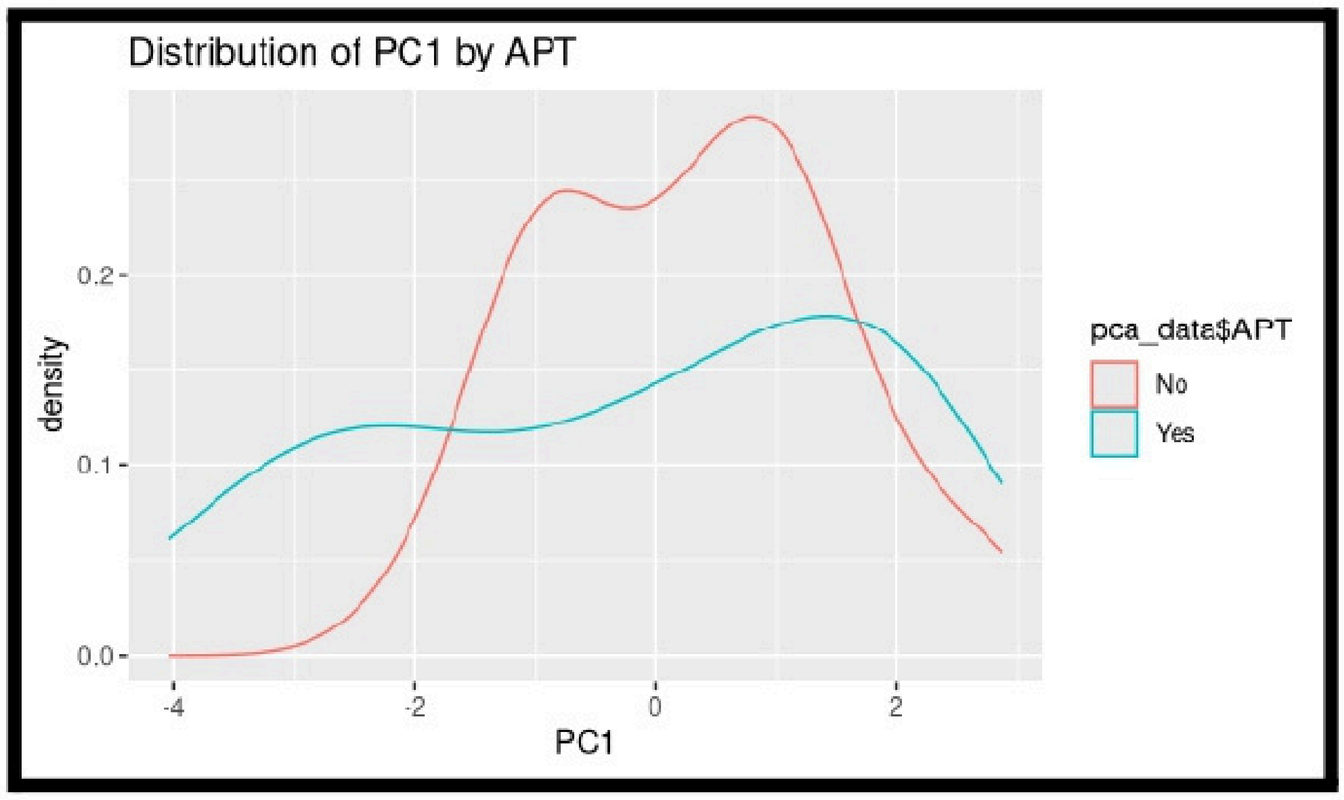
Thus, the PCA analysis reveals that features such as packet size and source IP address count are critical for distinguishing between APT and non-APT activity, while **IP.Range.Trust.Score** plays a role in identifying non-APT behavior. The PCA plot and loading vectors together help illustrate the underlying structure of the dataset, showing which features most strongly differentiate between APT and non-APT activities.

(vii) Which PC best assists in classifying APT?

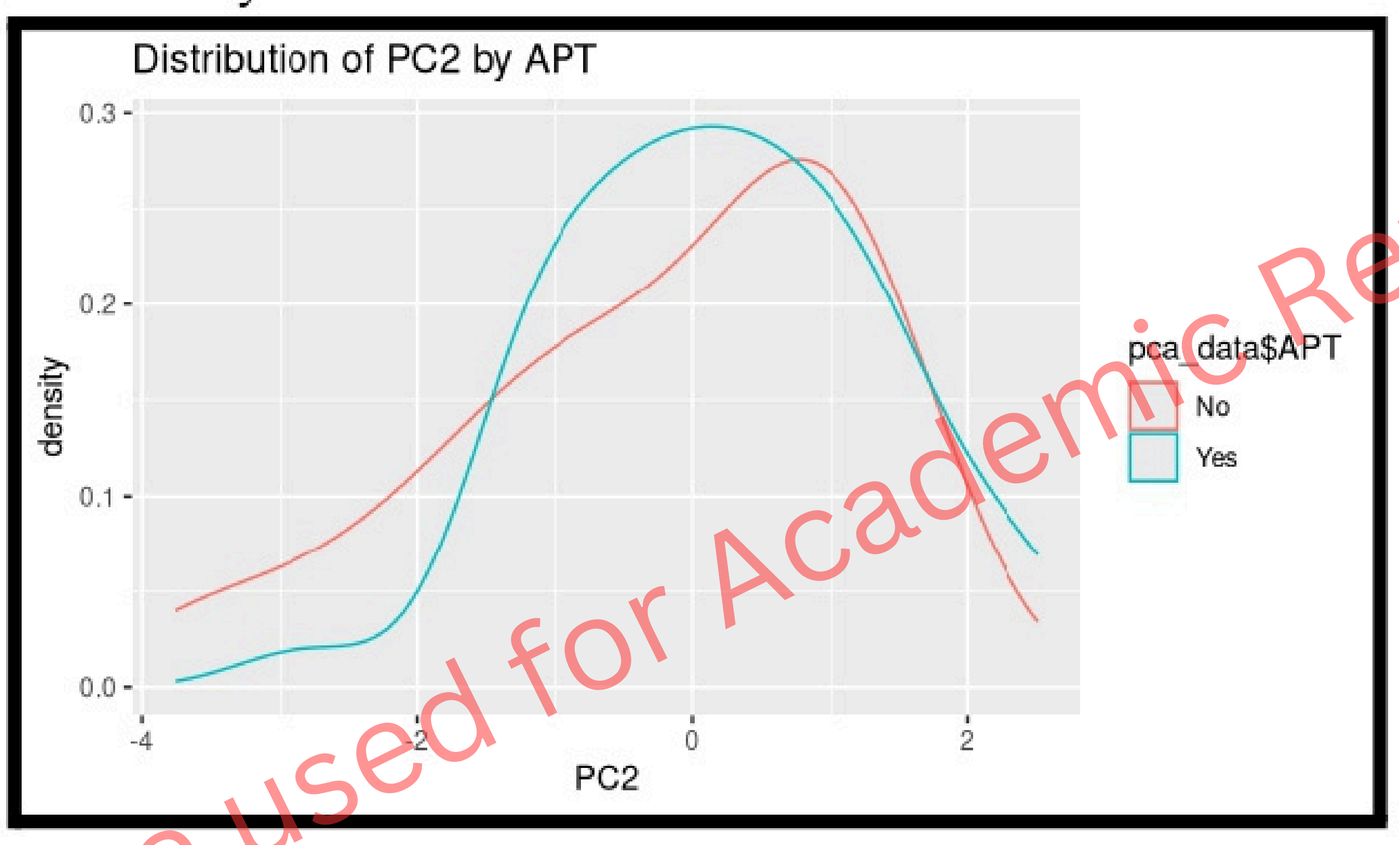
In the case of PC1, there is noticeable separation between APT= “Yes” and APT= “No”, despite some overlap. The distribution curves are distinctly shifted, indicating PC1 captures

"VAH Sample Question Library"

key variance related to APT behavior.



In the case of PC2, there is significant overlap between the two classes. The distribution for "Yes" and "No" is nearly identical. It suggests that PC2 does not effectively distinguish between the two categories. So, PC1 is the most effective principal component for classifying APT activity.



"To be used for Academic Reference Only"

Key Features Driving Classification from PC1 Loadings:

The features with significant contributions to **Principal Component 1 (PC1)** are those with absolute loadings greater than 0.3, indicating their strong influence on the direction of PC1. Based on the provided data, the following variables are the most impactful:

1. Average.Request.Size.Bytes – Positive Loading:

- **Interpretation**: A positive loading for **Average.Request.Size.Bytes** indicates that larger request sizes, which often occur in more prolonged or aggressive attacks, are strongly associated with APT (Advanced Persistent Threat) activity. Larger request sizes often correspond with larger data payloads, typical of cyber-attacks where substantial amounts of data are exchanged or transferred. This characteristic is often observed in more sophisticated, prolonged attacks, making it an important feature in distinguishing APT events.

2. Attack.Window.Seconds – Positive Loading:

- **Interpretation**: **Attack.Window.Seconds** refers to the duration of the attack window, and its positive loading indicates that longer attack windows are associated with APT activities. A longer attack window implies a more sustained and methodical attack pattern, characteristic of APT groups that conduct ongoing, persistent operations. This aligns with the idea that APTs are more gradual and stealthy compared to shorter, more explosive attack types.

3. Average.Attacker.Payload.Entropy.Bits – Positive Loading:

- **Interpretation**: **Average.Attacker.Payload.Entropy.Bits** measures the randomness or unpredictability of the payload in the attack. A high entropy value suggests that the payload is more complex and variable, which is often indicative of sophisticated attacks, such as APTs. High entropy can make it more difficult for traditional detection systems to recognize malicious behavior, as it may involve obfuscated or encrypted data. Therefore, this feature is a key indicator of APT activity, where attackers employ techniques to evade detection.

4. IP.Range.Trust.Score – Negative Loading:

- **Interpretation**: **IP.Range.Trust.Score** has a negative loading, meaning that lower trust scores in the IP range correlate with higher PC1 values, which are indicative of APT behaviors. A lower trust score suggests that the attack is coming from suspicious or untrusted IP ranges, which are often linked to malicious or hostile sources. This can point to IP addresses associated with botnets, proxies, or regions known for cyber-attacks, further confirming the presence of APT activity.

Combined Interpretation:

- **APT Activity**: The **positive loadings** of **Average.Request.Size.Bytes**, **Attack.Window.Seconds**, and **Average.Attacker.Payload.Entropy.Bits** suggest that APT attacks tend to be long-duration events that involve substantial data exchanges and complex payloads with high randomness. These characteristics are often associated with highly targeted, covert cyber-attacks.
- **Non-APT or Suspicious Behavior**: On the other hand, **IP.Range.Trust.Score** with a **negative loading** highlights that suspicious or untrusted IP ranges, which are often used by malicious actors, are also key indicators of APT activity. Lower trust scores reflect the involvement of IP addresses that are not typically associated with legitimate or safe activities, thus reinforcing the association with APT behavior.

"VAH Sample Question Library"