Hi class,

Below is the solution to Assignment 1. Please note that this is only for one random subset, so the numbers and figures will not be exactly the same as yours, but will be similar for the most part. **Please note the 'Data issues' below and that you will need to address them for Assignment 2.**

The biggest issue that I've seen from students' work in the idenfication of outliers is that many of you did 'not' use the tables to justify your decision as why there may be a problem. However, many students have just jumped to using boxplots to 'identify' outliers.

Task 1

Table 1 Summary of the categorical features

Feature	Category	N (%)
AlertCategory	Alert	53 (6.6%)
	Info	0 (0%)
	Informational	48 (6%)
	Warning	699 (87.4%)
NetworkEventType	NormalOperation	54 (6.8%)
	Policy_Violation	248 (31%)
	PolicyViolation	474 (59.2%)
	ThreatDetected	24 (3%)
NetworkInteractionType	Anomalous	358 (44.8%)
	Critical	32 (4%)
	Elevated	38 (4.8%)
	Regular	0 (0%)
	Suspicious	372 (46.5%)
	Unknown	0 (0%)
SessionIntegrityCheck	TRUE	398 (49.8%)
	FALSE	402 (50.2%)
ResourceUtilizationFlag	TRUE	670 (83.8%)
	FALSE	130 (16.2%)
Classification	Malicious	400 (50%)
	Normal	400 (50%)

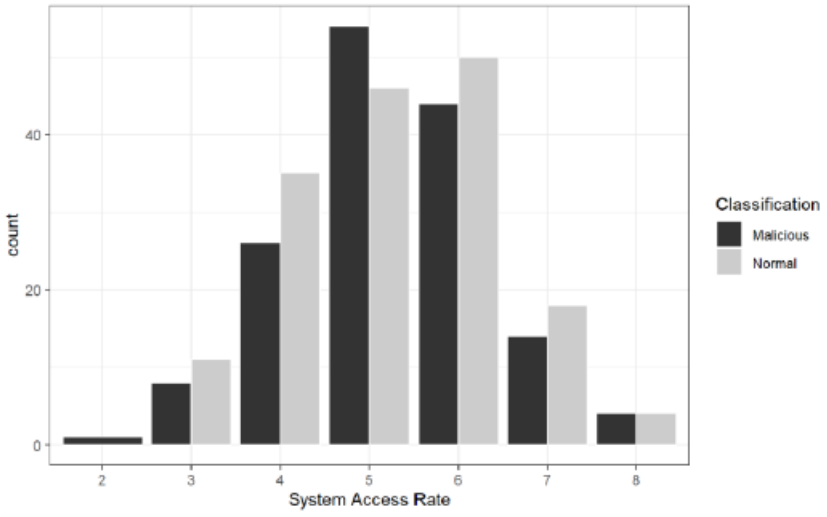
Table 2 Summary of the numeric features

Feature	Missing N (%)	Min	Max	Mean	SD	Median	Skewness
DTVI	0 (0%)	62249520.0	234905528.0	136501146.7	23151485.9	132596006.5	0.8
DTVO	0 (0%)	44152135.0	238631439.0	122101631.5	36483151.6	114027053.0	0.6
TPS	0 (0%)	15879.0	52283.0	29747.4	6504.0	28604.0	0.6
NAF	0 (0%)	-1.0	54205.0	31857.9	10314.7	33300.0	-1.4
UAL	0 (0%)	2.0	9.0	5.4	1.1	5.0	0.0
SAR	466 (58.3%)	2.0	9.0	5.2	1.2	5.0	0.2
SRL	0 (0%)	52732612.0	247364893.0	152860478.9	31106818.2	153087007.5	-0.1
RT	0 (0%)	9.1	99999.0	7653.2	26548.0	31.0	3.2

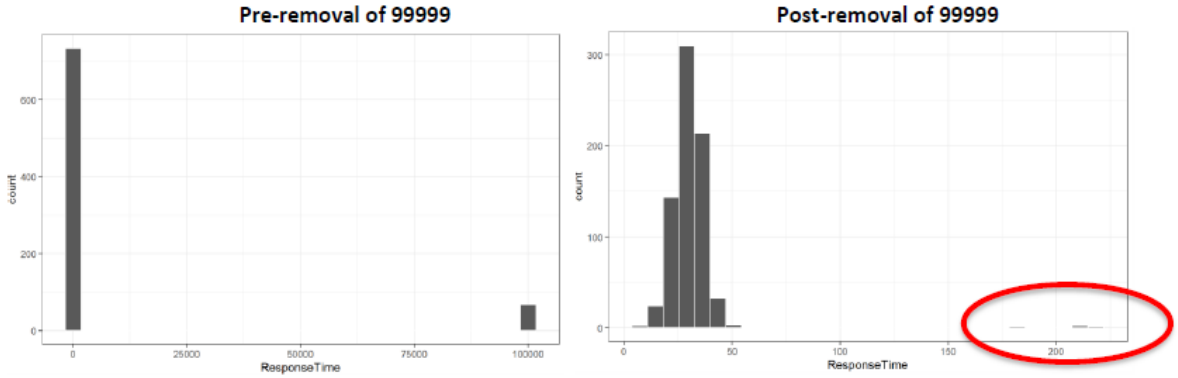
DTVI = DataTransferVolume_IN; DTVO = DataTransferVolume_OUT; TPS = TransactionsPerSession;
NAF = NetworkAccessFrequency; UAL = UserActivityLevel; SAR = SystemAccessRate; SRL = SecurityRiskLevel;
RT = ResponseTime

Data Issues:

a) Mis-labelled categories for AlertCategory and NetworkEventType. 'Info' and 'Informational' should be combined. Likewise with 'Policy_Violation' and 'PolicyViolation'.
b) Invalid data entry of -1 for NAF, which is a count variable and therefore should always be ≥ 0. Removing -1 will also reduce the skewness.
c) High proportion of missing values for SAR. Should be removed from further analysis. Furthermore, SAR is a weak discriminator of malicious samples (see image below).



d) High skewness (3.2) value for RT, indicating high likelihood of outliers and the 99999 entries are largely responsible for this (see 1st plot below).



However, there appears to be more outliers even after the removal of the '99999' entries (2nd plot). In this instance, and given that the main cluster is < 75, it is reasonable to remove the observations where the response time > 150 or one can use the +/- standard deviation since the main cluster resembles a normal distribution. Note: students are expected to provide the count and % of outliers in this instance.

Task 2

The data should be standardised in this instance since the continuous features have vastly different scales. Standardising will ensure each feature is treated equally in the PCA process.

Table 3 Variance explained by the first three principal components

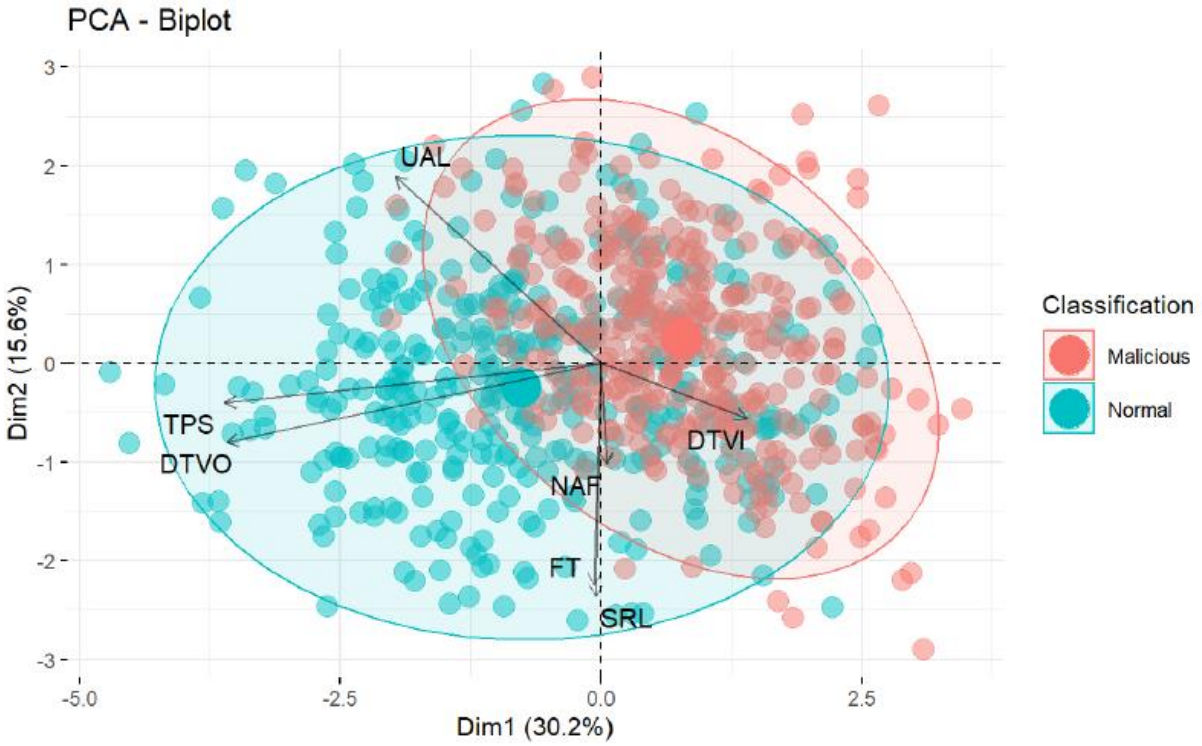
	PC1	PC2	PC3
Standard deviation	1.457	1.032	1.024
Proportion of Variance	0.303	0.152	0.150
Cumulative Proportion	0.303	0.455	0.605

Three PCs are required to explained at least 50% of the variability.

Table 4 Loadings for the first three principal components. Key drivers are highlighted.

Feature	PC1	PC2	PC3
DTVI	0.225	-0.042	0.658
DTVO	-0.634	0.124	0.166
TPS	-0.643	0.028	0.275
NAF	0.01	0.047	0.559
UAL	-0.364	-0.258	-0.345
SRL	0.033	0.656	0.013
RT	-0.015	0.695	-0.179

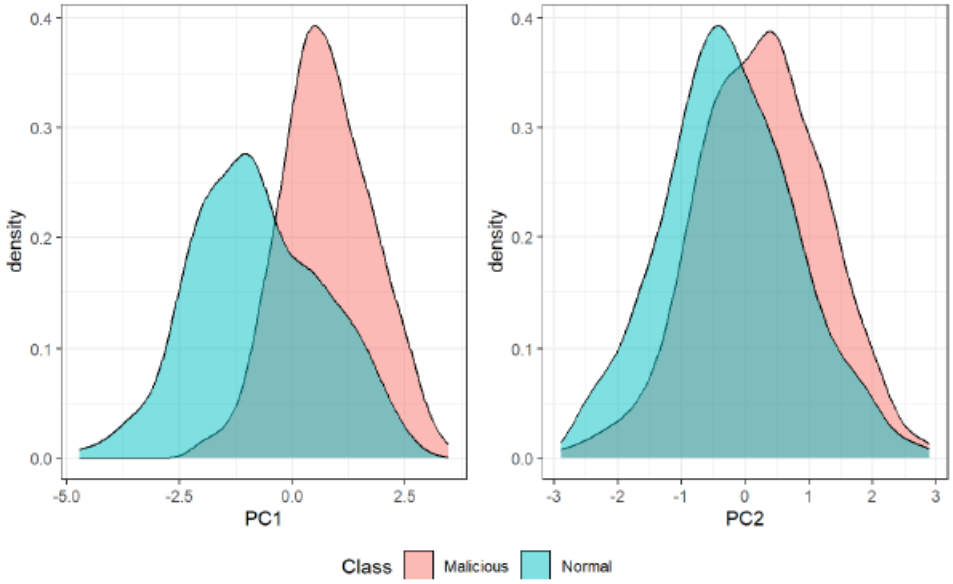
1) DTVO and TPS are the key drivers for PC1 as they highest absolute loadings (>0.6). Both are negatively correlated to PC1.
2) Similarly, SRL and RT are the key drivers for PC2 and both are positively correlated to PC2.
3) Same can be said for DTV1 and NAF with respect to PC3.



Interpretation:

1) PCA plot – Significant overlap between the two 'Malicious' and 'Normal' clusters on the right side of plot. However, samples are predominantly 'Normal' when PC1 < 1.
2) Loadings plot – TPS and DTVO are highly positively correlated to each other (angle <90

degrees), but both are weakly correlated to NAF, FT and SRL (angle ~ 90 degrees) and are negatively correlated to DTVI (angle >90 degrees). UAL is positively correlated to TPS and DTVO, but negatively correlated to the other features.
3) Biplot – Normal samples will tend to have higher TPS and DTVO (i.e. the vectors are pointing in the direction of the sample, i.e left), and to a lesser extent, higher value UAL but lower values DTVI (i.e. samples are located opposite to the direction of the vector).



The above figure shows the density plot of the samples when projected onto PC1 (left) and PC2 (right). From this figure it is fairly evident that PC1 would perform better in classifying samples as there is less overlap between their corresponding distributions. In particular, samples with PC1 < 1 are highly likely to be ‘Normal’ samples.