2024

Data Cleaning and Visualization

ASSIGNMENT 1

LOVYA BAJAJ

10565489

OFF-CAMPUS

Contents

Part 1 – Exploratory Data Analysis

Summary Tables

The following is a table representing the categorical variables and the number and percentage of each of their categories.

Categorical Feature	Category	N (%)
Alert Category	Informational	53 (6.6)
	Info	0 (0.0)
	Warning	705 (88.1)
	Alert	42 (5.3)
	Missing	0 (0.0)
Network Event Type	NormalOperation	64 (8.0)
	Policy_Violation	236 (29.5)
	PolicyViolation	470 (58.7)
	ThreatDetected	30 (3.7)
	Missing	0 (0.0)
Network Interaction Type	Regular	0 (0.0)
	Elevated	34 (4.2)
	Suspicious	367 (45.9)
	Anomalous	357 (44.6)
	Critical	41 (5.1)
	Unknown	1 (0.1)
	Missing	0 (0.0)
Session Integrity Check	True	393 (49.1)
	False	407 (50.9)
	Missing	0 (0.0)
Resource Utilization Flag	False	682 (85.3)
	True	118 (14.7)
	Missing	0 (0.0)
Classification	Malicious	400 (50.0)
	Normal	400 (50.0)
	Missing	0 (0.0)

Table 1.1: Number and percentage for Categorical Variables.

The following table summarises all numeric variables by calculating their min, max, men, median, skewness and missing values.

Continuous Feature	N (%) Missing	Min	Max	Mean	Median	Skewness
Dala Transfer Volume IN	0 (0.0)	35195972	217799173	136721648	133042056	0.1
Dala Transfer Volume OUT	0 (0.0)	43581921	239367593	121476291	114059148	0.6

Transactions Per Session	0 (0.0)	11414	52363	29606.9	28831.5	0.6
Network Access Frequency	0 (0.0)	-1	54790	31935.45	33284	-1.4
User Activity Level	0 (0.0)	2	10	5.4	5	0.3
System Access Rate	469 (58.6)	3	8	5.3	5	-0.1
Security Risk Level	0 (0.0)	63662913	238717831	150679902	150413214	0.0
Response Time	0 (0.0)	7.5	99999	7901.9	29.9	3.1

Table 1.2: Summary for Numeric values.

List of Data Issues

Yes, there are 2 invalid categories in the categorical table. In Alert Category feature, there is "Info" heading which seems to be same as "Informational" and is misleading and gives error in the total Informational percentage. Similarly, in Network Event Type, "PolicyViolation" and "Policy_Violation" is a mistake when they represent the same category.

Yes, there are outliers as can be observed through the table and the percentages above.

Network Access Frequency: Outlier is -1, since it is very odd from other values and entirely different from the mean. It is 44 in number and approx. 5.5% of total values in Network Access Frequency.

Response Time: 99999.0 is an outlier since it is comparatively a very high number than the rest of the values as well as the median and is 63 in number and 7.88% of all values in Response time.

Part 2 – PCA and Data Visualization

PCA and Results Interpreted

The data needs to be scaled/standardised as the values of different categories varies within a huge range. So, it is needed to be standardised so that the large range values do not undermine the smaller range values.

Individual and Cumulative Proportions of Variance

	PC1	**PC2**	**PC3**
Proportion of Variance	0.268	0.150	0.139
Cumulative Proportion	0.268	0.418	0.558

Looking at the data in the table above, 3 of the principal components (PCs) are adequate to explain at least 50% of the variability. As if we take 1 PC, it would be approx. 26% and with 2 PCs, it would be approx. 41% but with 3 PCs, 55.8% of variability can be explained.

Coefficients for PC1, PC2, PC3

	PC1	PC2	PC3
Data Transfer Volume IN	0.244	-0.378	0.622
Data Transfer Volume OUT	-0.622	0.066	0.270
Transactions Per Session	-0.625	-0.090	0.345
Network Access Frequency	-0.133	-0.067	-0.366
User Activity Level	-0.366	-0.031	-0.453
System Access Rate	-0.093	-0.540	-0.079
Security Risk Level	0.032	0.646	0.273
Response Time	0.042	-0.360	0.028

For PC1, the key drivers are Data Transfer Volume OUT and Transactions Per Session having the highest negative value. Similarly, for PC2, key driver is the Security Risk Level with a coefficient value of 0.646, and for PC3, it is Data Transfer Volume IN since it has the highest positive value than all other features.

PC1 vs PC2 Biplot

The following is a biplot image of PC1 vs PC2 which represents PCA results and helps in visualising them.

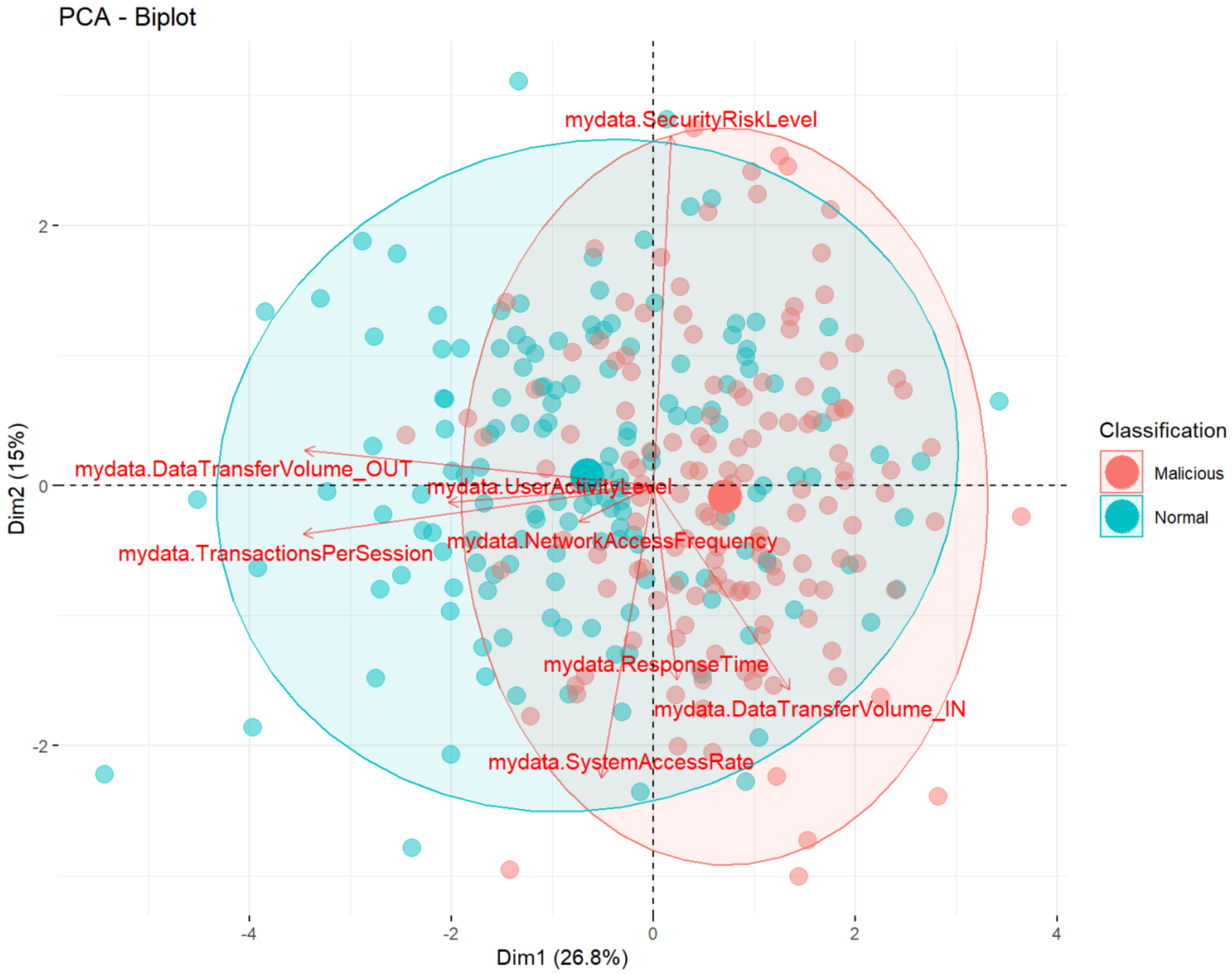

Image 2.1: Biplot for PC1 vs PC2.

PCA Plot

Looking at the figure above along with Dim1 i.e. PC1 axis, red dots representing Malicious events and blue dots representing Normal events are overlapping and are both present in groups mainly between approximately -1 and 2. So, it can be said that there is presence of clustering and grouping between malicious and normal events.

Loading Plot

Security risk level feature is the most important of all since it is the longest vector and it is most important to PC2 as it is most parallel to it. Other than this, Data Transfer Volume OUT and Transactions per second are very important to PC1.

By observing the angles between the two vectors, the following relations can be interpreted:

- Data Transfer Volume OUT, User Activity Level, Network Access Frequency, and Transactions per Session are positively correlated with one another since they make less than 90-degree angle between their vectors. Similarly, System Access Rate, Response Time, and Data Transfer Volume IN are positively correlated with one another.
- Data Transfer Volume OUT is uncorrelated with System Access Rate since the angle between them is approximately 90 degrees and negatively correlated with Data Response Time and Data Transfer Volume IN since they make an angle greater than 90 degrees.
- User Activity Level is positively correlated with System Access Rate and uncorrelated with Response Time and negatively correlated with Data Transfer Volume IN.
- Transactions Per Session is uncorrelated with Data Transfer Volume IN.
- Network Access Frequency is almost not related with Response Time and Negatively Correlated with Data Transfer Volume IN.
- All the vectors are mostly negatively correlated with Security Risk Level except with Data Transfer Volume OUT which is positively correlated with Security Risk Level but not to a large extent.

PCA Plot + Loading Plot

After analysing both PCA plot and loading plots together, it can be interpreted that Normal Event Types tend to have higher values of Data Transfer Volume OUT, User Activity Level, and Transactions Per Session whereas Malicious events tend to have greater values of Data Transfer Volume IN, Security Risk Level and Response Time and lower values of Data Transfer OUT, User Activity Level, Transactions Per Session and Network Access Frequency. This can be said by analysing the direction of the vectors as well as the presence of malicious and normal event types together. It can help us identify between the malicious and normal events.

Selection of Dimension

PC1 can be used to identify the malicious events more than PC2 since there is a good separation between the points of Normal and Malicious events along with PC1 axis rather than with PC2 axis. This can be seen by projecting PCA points along the PC1 axis and then along PC2 axis. In simple words, PCA points can be plotted along PC1 by keeping PC2 coordinates 0 and vice versa.

The following images, Image 2.2 and Image 2.3 shows the difference between the separation between malicious events and normal events when PCA points are projected along PC1 axis and PC2 axis respectively.

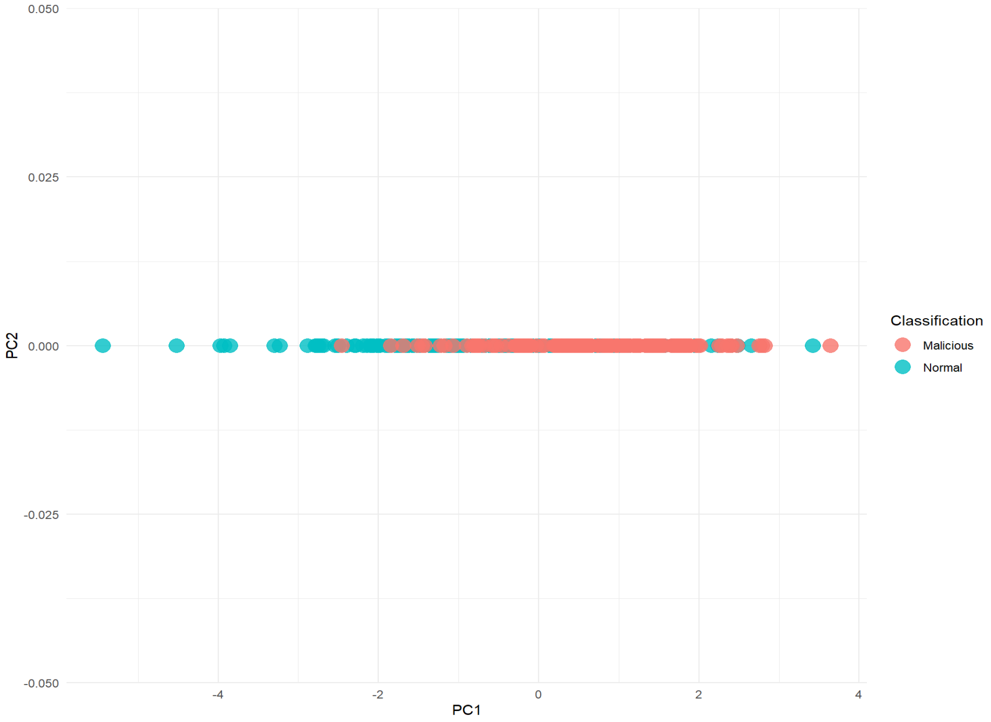

Image 2.2: Projection of points on PC1 axis

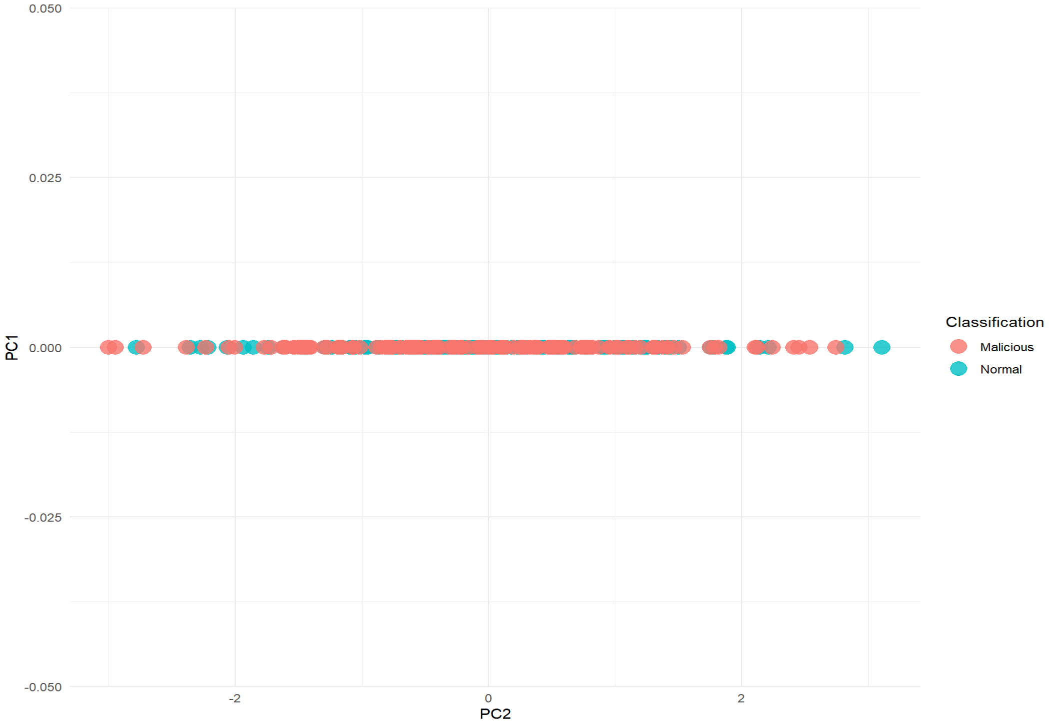

Image 2.3: Projection of points on PC2 axis.

REFERENCES

Zach. (2021, June 22). *How to Use sum() Function in R (With Examples)*. Statology. https://www.statology.org/sum-function-in-r/

Zach. (2021a, June 18). *How to Use is.na in R (With Examples)*. Statology. https://www.statology.org/is-na/