Assignment 1 Part B – PCA and Visualisation

A. Data Preparation

The data set provided had a few issues that needed to be cleaned up before any principal component analysis could take place. The first issue was addressing the missing values in the 'Outside Network' and 'Verified as Malware' columns. ***
****

Another issue was converting the categorical variables of the data set into something usable for the principal component analysis. This was achieved using dummy variables in which categorical variables with values of 'Yes' were represented as a 1, and categorical variables with values of 'No' were represented as a 0. The 'dummy.data.frame.' function is a part of the 'dummies' package, and it was used to transform the categorical variables into dummy variables. It does this by adding an extra column for each variable. For example, 'outside.network' was split into two new columns: 'outside.networkYes' and 'outside.networkNo' In the 'outside.networkYes' column all 'Yes' values are now represented as a 1, and all 'No' values are represented as 0. Conversely in the 'outside.networkNo' column, all 'No' values are represented as 1, and 'Yes' values are represented as 0. The 'Verified as Malware' variable was converted back into a factor that contained 'Yes' and 'No' values, using the cut column and was rebound to the data set. Lastly, ***********************
**************. The data set was now ready to be used for principal component analysis.

B. Proportions of Variance / Coefficients

Importance of components:

	PC1	PC2	PC3
Standard deviation	1.4570	1.0983	1.0750
Proportion of Variance	0.2359	0.1340	0.1284
Cumulative Proportion	0.2359	0.3699	0.4983

Commented [JL1]: Do not copy and paste from R. Tabulate in Excel.

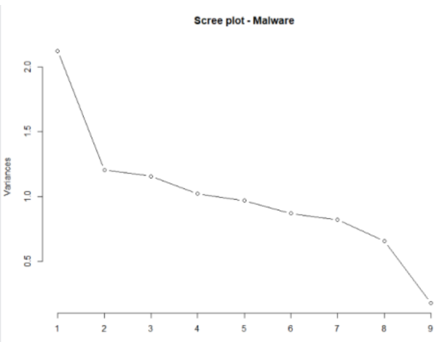
Upon performing principal component analysis of the data set, a summary was performed that revealed the standard deviation, the proportion of variance, and the cumulative variation of the nine principal components. This revealed that PC1 explains ********************** PC2 explains ************, and PC3 *********** Cumulatively ***************************************
**********************************.

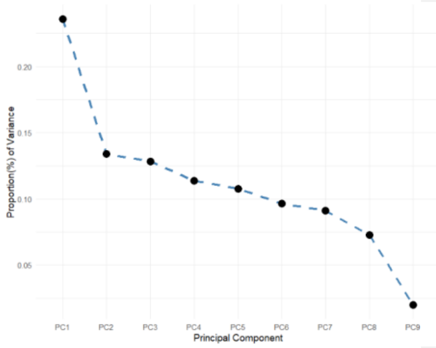
	PC1	PC2	PC3
Num.attachments	0.62565596	-0.05846506	-0.1037421
inc.executable	0.09767587	0.16185373	-0.6587857
inc.ZIP	-0.13224594	-0.26748208	-0.2691932
inc.PDF	0.37395007	0.30869970	-0.1195401
inc.DOC	0.28146353	0.40142270	-0.1014450
Unknown.Format	0.49317220	-0.46312829	0.1364324
URL.count	0.00215855	-0.63331623	-0.1391389
Outside.Network	-0.08258719	-0.14541717	-0.6324444
Email.Size	0.33529176	-0.07536208	0.1422451

Commented [JL2]: Same coment as above.

The above image shows the coefficients for PC1 to PC3. By looking at this data, it is obvious that the 'num.attachments' is most important to PC1 ******************. PC2 is defined by the 'URL.count,' ****************** Lastly, PC3 is defined by two features as they are both of very similar values, **.

C. Scree Plot

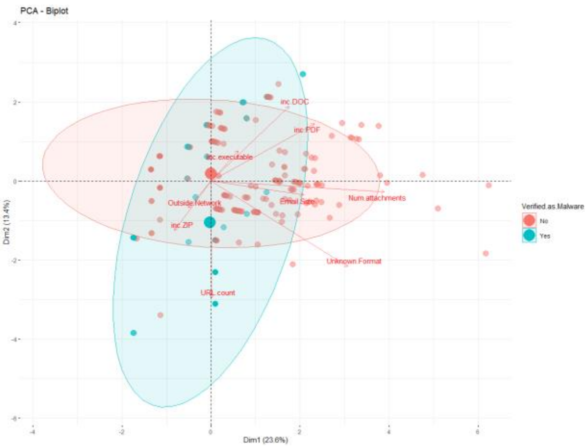




A scree plot was created to visualize and determine how many principal components should be retained. The scree plot shows the proportions of variance for each principal component. ********** ** ***.

Commented [JL3]: Need a bit more her about the trends in the scree plot.

D. Biplot



The PCA biplot shows that there is quite a bit of overlapping of samples that were verified as malware and samples that were not. *** ** ** ****************.

The 'incDOC', 'incPDF', and 'inc.executable' vectors ************************ with one another and highly are *******************'Outside.Network, 'inc.ZIP' and 'URL.count'.

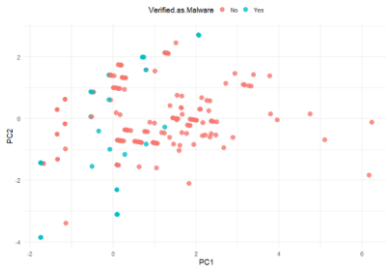
‘Outside.Network’ and ‘inc.ZIP’ are ********** and have some correlation with ********** and ‘Num.attachments’ are highly and positively ********** ********** Although they are not negatively correlated with all remaining vectors, they have no correlation with them either.

The biplot shows that samples that are not verified as malware, in general, have ********** ********** It also shows that these samples more often ********** **********. Samples that were malware had a higher ********** and a larger number of samples that **********.

E. Data Classification

The data set provided had very little separation between samples that were verified and samples that were not on both PC1 and PC2. Therefore, it is very difficult to choose either dimension to assist with classification. However, if having to choose between the two, ********** ********** **********. The key features in this dimensions that drive this process are:

****
****
****
****



F. Cross-Tablatures

****

		Verified as Malware	
		Yes	**No**
**********	**Yes**	26 (49%)	110 (37%)
	No	27 (51%)	187 (63%)
	Total	53	297

26 of the samples that were verified as malware were of **********, and 27 were of a known format. 110 samples that were not verified as malware were of **********, and 187 samples were of a **********. Samples that were verified as malware had a near fifty-fifty split of being an **** ********** **********

****

		Verified as Malware	
		Yes	**No**
**********	**Yes**	7 (13.2%)	67 (22.5%)
	No	46 (86.8%)	230 (77.5%)
	Total	53	297

7 of the samples that were verified as malware included a PDF file, and 46 of them did not include one. 67 of the samples that were not verified as malware included a PDF, and 230 samples did not. This indicates that the majority of verified malware did not contain a PDF, and the same goes for samples that were not verified.

Commented [JL4]: The raw numbers are not too meaningful here given the discrepancy in the number of Yes and No cases. it's more about the %.

****

	Verified as Malware	
	Yes	**No**
**********	1 (0)	1 (2)

The samples that were verified as malware and those that were not both had a median of 1. The IQR of samples that were malware was 0, indicating little variation in these samples. A boxplot was used to further visualize these samples, which revealed that all samples that were verified as malware had at least one attachment. Samples that were not verified as malware had an IQR of 2, indicating more variation in its samples compared to samples that were verified, although this was still fairly minimal.

****

	Verified as Malware	
	Yes	**No**
**********	6217 (28441)	25306 (214455)

The ********** was 6217, with an interquartile range of 28441. The variability of 'Yes' samples was tame compared to samples that were not verified as malware. The median of samples that were not verified as malware was 25306, with an interquartile range of 214455. This indicates that 'No' samples had ********** **********.

G. Data Issues

Commented [JL5]: Missed a few.

Commented [JL6]: This is not an issue.

There were two issues with the data set that made the principal component analysis difficult. The first issue was the number of variables in the data set that was to be used in the principal component analysis. Having nine variables meant that there would be nine principal components. This resulted in losing a large amount of the variability of the data set when it came to the analysis. Realistically the first two principal components are desired to be used for PCA with the aim of retaining most of the variability. In this case, even if the first four components of the data set were to be retained, it would still only be a cumulative variance of ********** Another issue with the data set was the lack of variation between samples that were verified as malware and samples that were not. ********** **********. Without ********** it would be difficult to make any assumptions on any future samples. In conclusion, these issues made PCA difficult, and therefore, the data set was inappropriate for **********.