1. **Data cleaning**
2. **Filling NA Values**

In order to match the NA values in the columns outside.network, and is.malware, first, the dataset needs to be standardized. For this purpose, strings as factors have been changed to false, to make the re-encoding of strings without any issues (John, 2018). Once the value has been changed the missing values in both the columns replaced with the variable No. The reason behind this is, in both the columns the presented values stated as Yes, so thus the No has been used to replace the NA values in both the columns. For this a simple is.na function (Joachim, 2020) has been used.

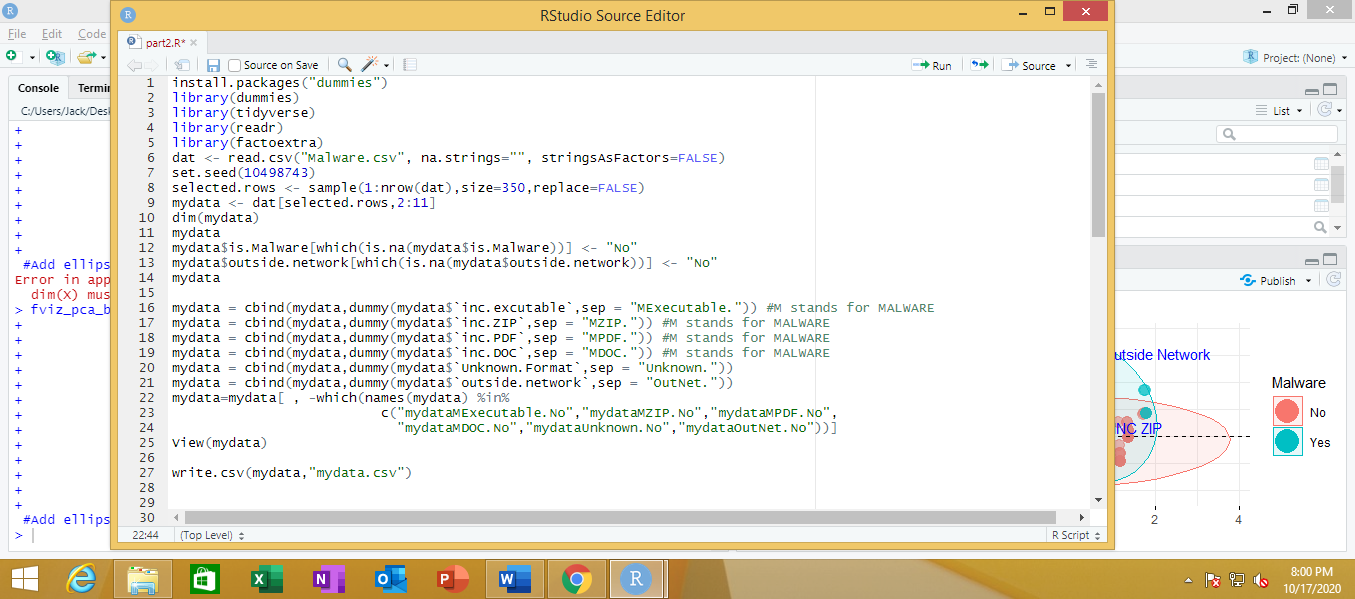


Figure Code to fill NA values with No

1. **Converting category variable**

Since the values inside the dataset are strings and PCA requires numeric values, need of conversting category variable to numeric values has to be done. For this purpose dummy variable has been created for each column using generic cbind function. This cbind function basically used to merge multiple data frames (Olivia, 2020) By using this function each column has been specified individually. Once the cbind function is used it has created more new columns with the integer values. And once the dummy values created in new columns the cleaned dataset exported and assigned as mydata.

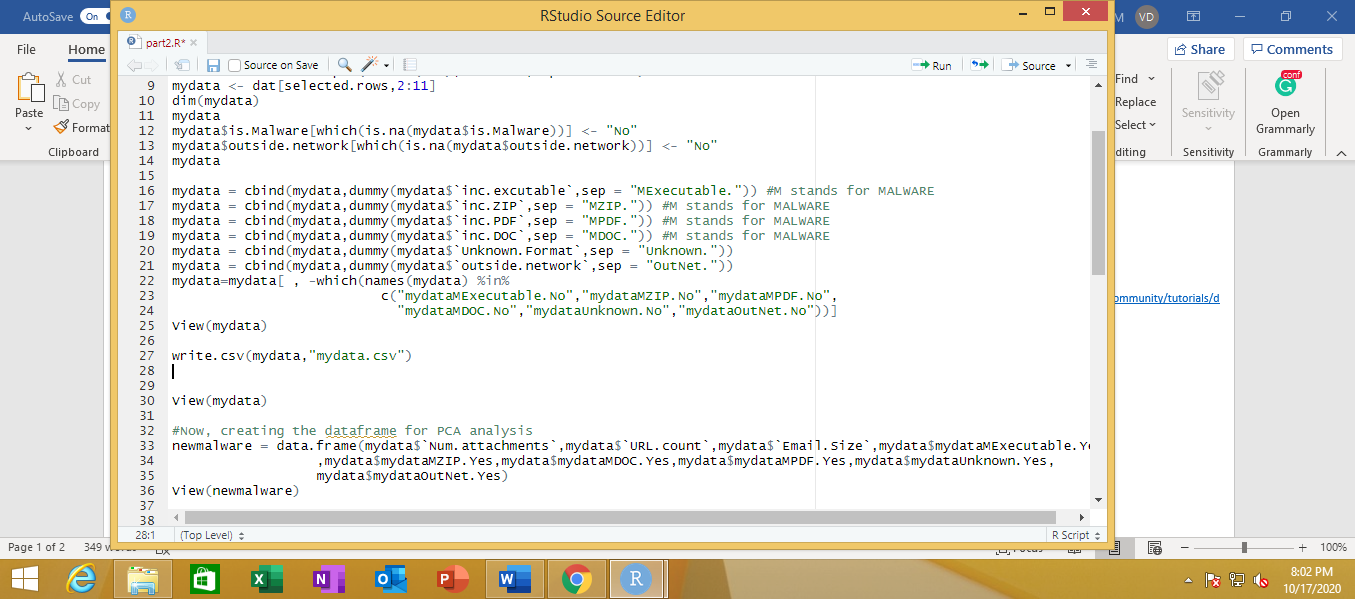


Figure Filling dummy values using the function cbind

1. **PCA**

To perform PCA, all the columns are kept inside the variable called newmalware. Once, the columns are inside PCA has been performed, the PCA summary of 9 features has been given in the following screenshot. From the screenshot, we can find out that the proportion of PC 1 higher than remaining PCs (in this case from PC1 to PC9). Narrowing to, from PC1 to PC4, the variation remains > 0.09 which can be considered as low moderate level difference.

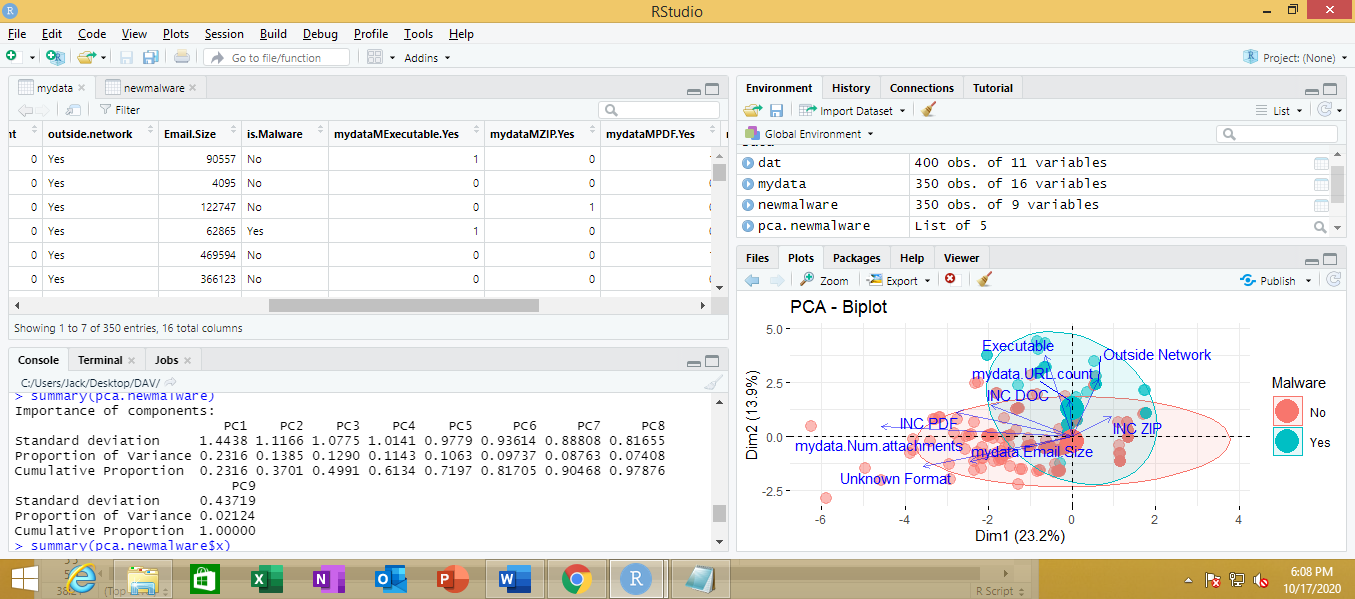


Figure 3 PCA Summary

The following screenshot shows the co-efficient value between from PC1 to PC4, from that we can see that the values derived PC1 and PC2 values of the features are completely vary. The PC1 to PC4 coefficient is used to show the first four variable matrix, by using the rotation function the eigenvectors has been derived here. The variables and principle components are postivily cor-related with the co-efficient with the postivie value, and vice-versa the negative correlation represented by negative conefficient values.

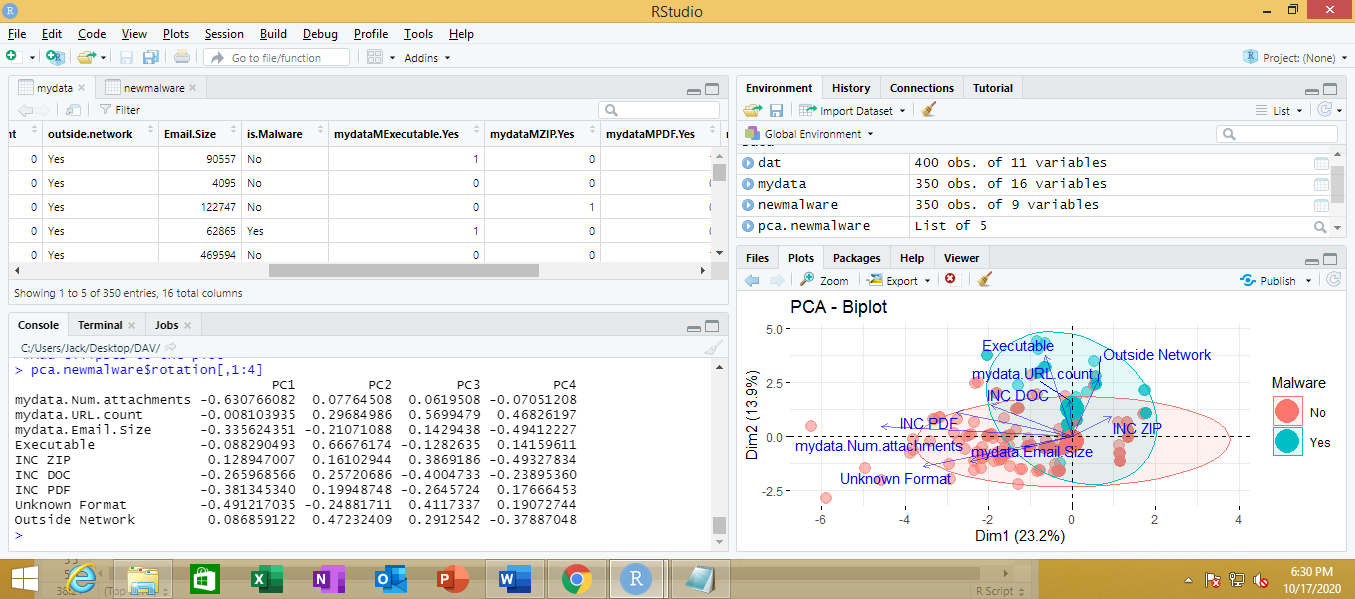


Figure 4 PC 1 to PC 4

In order to find out the number of principle components for minimum of 50% of data variability, proporation of variance has been done using a function called importance. This result explains that 3 PC is required to explain a minimum of 50% of the variability in this case. The following screenshot show the result of proporation variation values using the function importance.

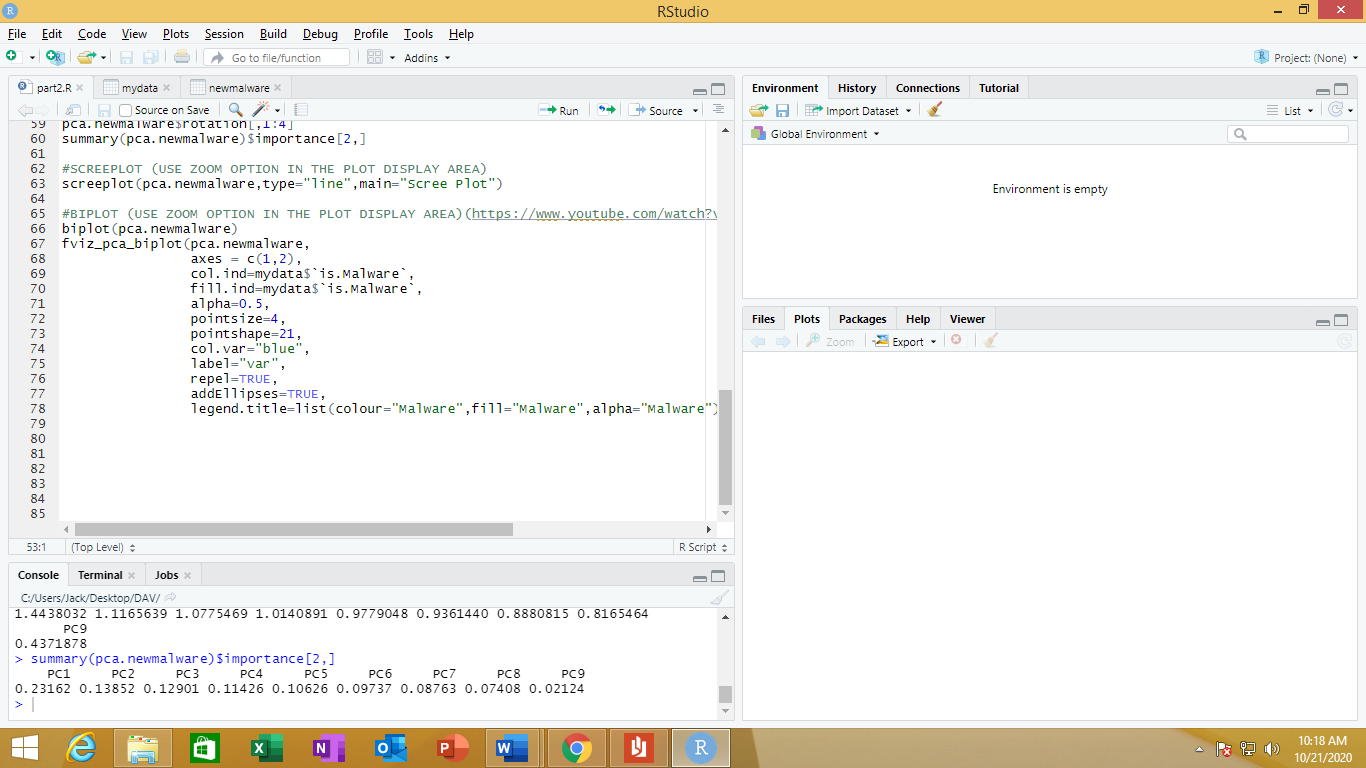


Figure proporation of variance

1. **Scree plot**

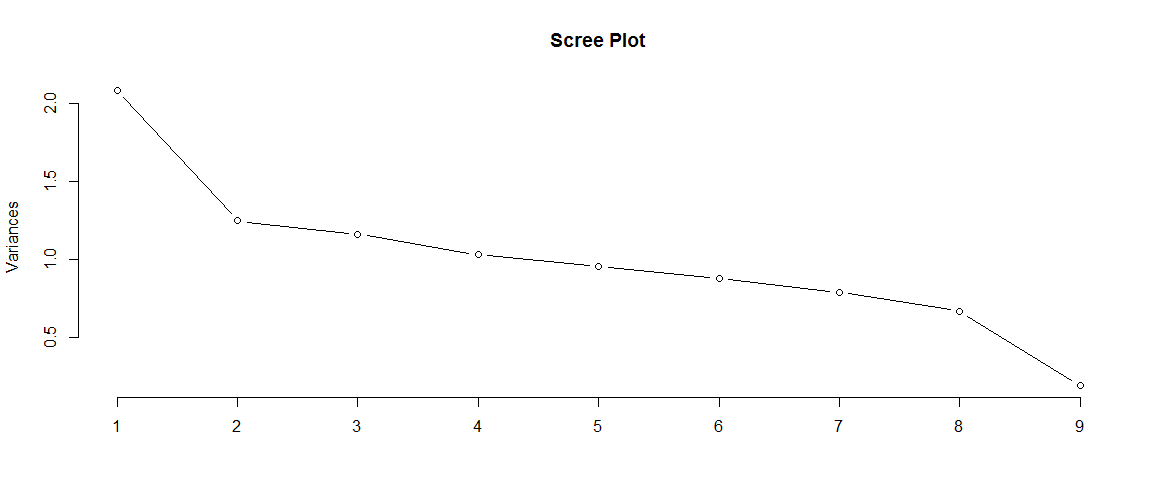


Figure 6 Scree Plot

The above screenshot shows the variation between the all the 9 features, in that it’s clear that features 1-4 are enough to summarize almost 80% of data.

1. **Biplot**

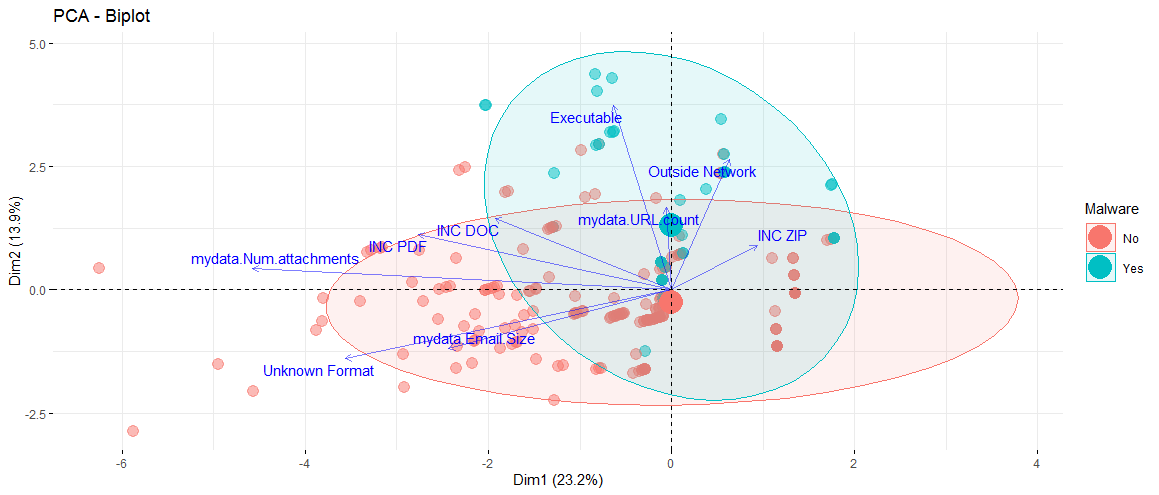


Figure 7 BI Plot

The above image shows the biplot created for the PCA. When comes to the dimensions PC1 dimenstions are repreneted in upper and bottom layer (Y axes), and the two sides right and left reprenstes the PC 2 (X axis) Through this biplot, we can see that the characteristic Executable has the highest length which indicates the importance of it. Also, we can see that the characteristics outside the network and ZIP are forming an angle which is representing their positive co-relation. On the other hand, the characters, executables and unknown format are diverging and indicates they are negatively correlated. Also, from this, we can see that the characteristics ZIP, outside network, Executable, URL count, Doc has higher values than the other characteristics.

**Selection and contribution**

Overall, the vertical axes of PC2 provide good separation between the classification, because the PC1 values are mostly pointing to towards non-malware and the lacks the proper separation. Whereas in PC2 the separation between the malware and non-malware indicates the important characteristic. Another important reason to choose PC2 is it’s positive correlation values. compairng with PC1 or any other the PC’s, variables and principle components of PC2 are more postivily correlated with the co-efficient

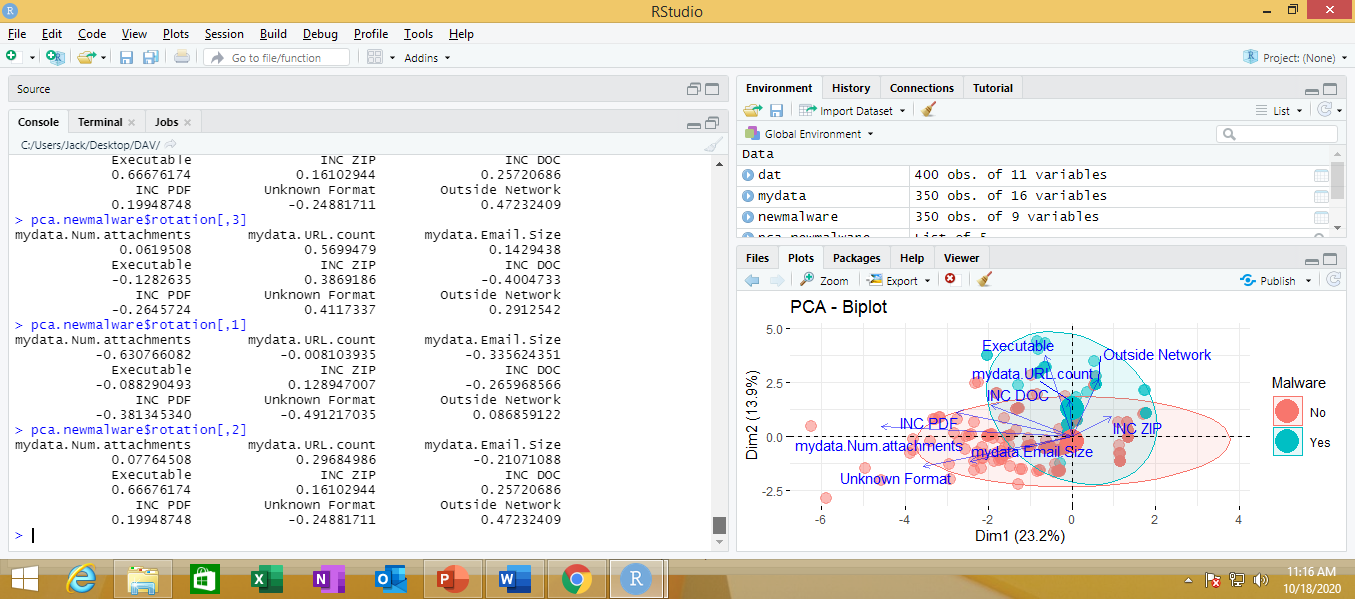


Figure 8 PC2 loadings

The above image shows the loading values of chosen PC 2, through this, we can see that the characteristics Executables, and outside network and followed by URL count, are the key features derived from the vertical dimension.

**References**

Joachim, S. (2020). *R is.na Function Example (remove, replace, count, if else, is not NA)*. Statistics Globe. Retrieved from https://statisticsglobe.com/r-is-na-function/.

John, M. (2018). R Tip: Use stringsAsFactors = FALSE [Blog]. Retrieved from https://www.r-bloggers.com/2018/03/r-tip-use-stringsasfactors-false/.

Olivia, S. (2020). *(Tutorial) Data Reshaping in R*. DataCamp Community. Retrieved from https://www.datacamp.com/community/tutorials/data-reshaping-in-r.

Paul. (2015). Replacing NA values with different values in Data Frames in R [Video]. Retrieved from https://youtu.be/w1t1a\_doQoo.

Jacques, W. (2017). Creating a dummy variable using R [Video]. Retrieved from https://www.youtube.com/watch?v=Qi9notslCzM.

Bharatendra,R (2017). Principal Component Analysis in R: Example with Predictive Model & Biplot Interpretation [Video]. Retrived from https://www.youtube.com/watch?v=OowGKNgdowA