RenderingTesting

Introduction

Multiple institutions have been a target of increasingly more disruptive or destructive cyber attacks over the last few years which has lead to government action.

The data used in my work was collected yearly by the uk government department for Digital, Culture, Media and Sport (DCMS) with the purpose of helping the government understand the importance cyber security for British institutions, better shape policy regarding cyber security, create schemes to increase awareness for such problems and better protect institutions form cyber security threats.

As such this analysis will investigate the relationship between how institutions protect themselves from cyber attacks and the affect of said attacks on these institutions in the last 5 years.

Objectives

Objectives of this report:

-Creating of a new tidy data set for each of the years including recompiled variables for the management, policing and rules implemented to protect the organisation, the type of attack that affected the institution and its respective outcomes.

-Utilizing Multiple Imputation by Chained Equations (Mice), to replace the missing data.

-Do hypothesis testing on my new fitted models to compare how the protective measures and attack type affect the outcome of the cyber attack on the institution.

-Mention the limitations of this analysis.

-Conclusion with recommendation for future research.

Data

Initial data wrangling:

Due to the untidy state of the data collect via the random probability telephone survey, these data sets containing between 421 to 462 variables have to been clean up into 21 easily comparable variables.

The clean up process consisted of computing new variables utilizing the multiple subcategories of answers to the survey questions, grouping them into more flexible options while adjusting missing values to allow for such computation maintaining the original binary design and increasing the scale of the size variable to produce better grouping and latter on better imputations due to the data sets didn't had the distinction between the intervals [250, 999] and [1000, ∞] that was present in the survey.

I also had to remove a few results from each year data set because these institutions still had their systems down after being attacks and since I don't have the information of the data of the attack and the data of the survey for those particular institutions it is impossible for me to quantify the time for restoring their systems, creating this way data that doesn't give us any possible information about the topic but is not missing, so it should not be replaced with missing data for computation.

Removing unnecessary data variables that are irrelevant to my hypothesis testing, converting the SPSS labelled data into R data structured factors and numerical to enable imputation of missing values and proper correlations computation and conversion of the appropriate missing value responses to actual missing data.

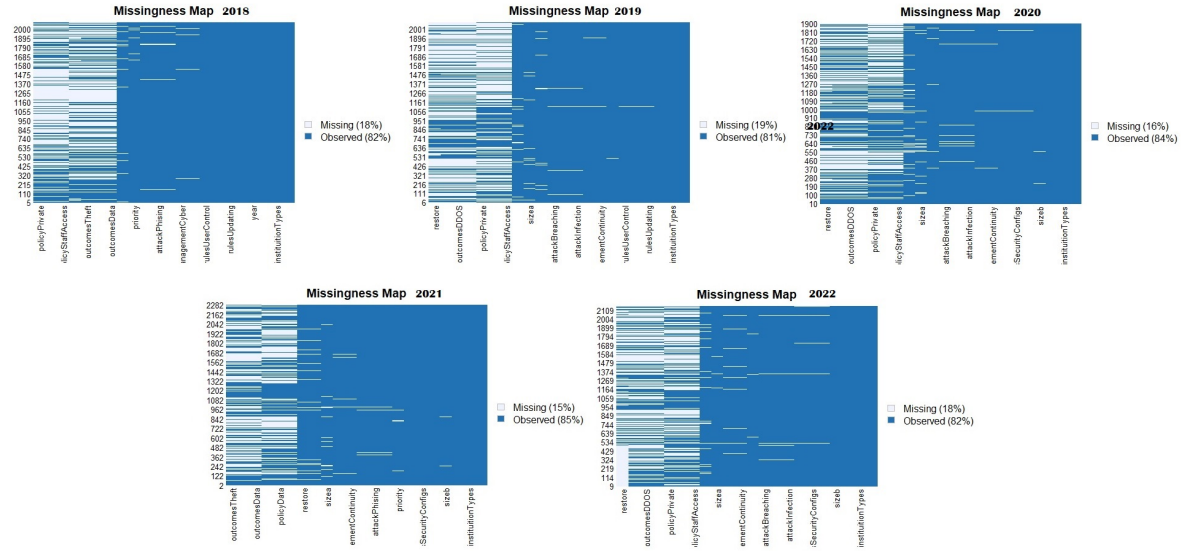

Figure 1: MissingnessMap

Data analysis:

It can be observed a significant degree of missing data on my data sets coming from multiple sources, the main source of missing data is derived from a limitation of the data collected, the lack of reporting channels in institutions leads to the majority of the missing values that lead to direct missing data in the data sets collected and indirect missing data by institutions answering that they do not know the answer to some of the questions in the survey. Lastly there is missing values associated to the type of attacks by institutions refusing to answer the question regarding the type of attack inflicted in their respective institution.

The missing data will have to be imputed using Multiple Imputation by Chained Equations on each of the data sets, for this I will be using the R library mice created by professor Stef Van Buuren.

For the imputation we had the consider the 3 following parameters, number of imputations, number of iterations per imputation and method for imputation.

The number of imputations was chosen following two rules, the first one is Relative Efficiency (RE) is lower with a higher number of imputations according to Rubin's formula RE=1/(1+(FMI/m)), where FMI is approximately equal to the percentage of missing data and m the number of missing data. [@rubin_1975] The second was a rule of thumb described in the book "Multiple imputation using chained equations: Issues and guidance for practice" where they recommended to equate the number of imputations to the percentage of missing data in each of the data sets which is what I will be using. [@White2010]

The number of iterations was chosen based on the convergence, that is when plotting the imputations the variance between the imputation chains is close to the variance of the chained imputations which is an indicator of an healthy convergence, this convergion was achieved after multiple trials with different numbers of iteration. [@AppliedMissingDataAnalysis]

TODO SHOW PLOT GRAPHS

Lastly for the method of imputation I choose not to use the default method ppm which is more appropriate for continuous data, most of the variables were imputed with the method of logical regression "logreg" due to the nature of the majority of the values being dichotomous binary variables, the numerical variable was instead imputed with the method of polynomial regression "polyreg" because size has a discrete finite number of values. [@AppliedMissing-DataAnalysis]

Results

##Missing data

##Visual analyses

##hypothesis testing

Limitations

There are multiple limitations to my analysis to be noted. Firstly, the data collected is limited to cyber attacks that were detected, there is variety of attacks that have gone unnoticed and therefore the data has a systematic tendency to underestimate the real level of breach attacks.[@CyberSurvey2020]

Secondly the missing data generated by imputation is biased since not all data is missing completely at random, mainly due to smaller and less staffed institution not having IT professionals and as such they don't have the infrastructure to detect, assess and report cyber attacks.

Conclusion and recomendations