RenderingTesting

Table of contents

Introduction

##talk about the importance of hacking, percentage of data loss link it with the data collected

stats about data breach

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.

The data collected contains information detailing the attacked institutions, the countermeasures in placed before and after the attack, the type of attack and it's affects on the company.

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 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 size of an institution will affect the time needed to restore business operations.

-Mention the limitations of this analysis.

-Conclusion with recommendation for future research.

Report structure

This report will be structure in the following order, firstly I will be describing the data set in more detail and my tidying process, secondly I will talk about my methodology for data analysis, afterwards in my results I will be displaying a visual analysis of the data, test results and it's meaning, afterwards I will discuss the limitations of my data and lastly conclude discussing the implications of my results for future research and the industry.

Data set

The data sets contain the data used for the statistical analysis done by the uk government DCMS department, they were collected and published in the uk data service, however they have not been made completely public and require a request and its approval to obtain access to each of the data sets.

Each one of these data sets contain the data associated to institutions affected by cyber attack with its multitude of implications for costs, business downtime, reporting and outcome as well as a detailed description of the policies, rules and investment in security measures to counter

such security threats and some key parameters to describe the institution such as size, market sector and better contextualize the 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.

There was also a further cleaning of the data sets by removing variables that were unused and not relevant to my my hypothesis and its associated testing

The data sets were previously compiled and run in SPSS which is a statistical software developed by IBM for data analysis, therefore all the data in the data sets were in SPSS data structures that needed to be converted to R structures such as numeric and factor to allow for imputation and model fitting.

Methodology

The process of the methodology will be starting with a simpler hypothesis test based on mean comparison to discovers the relationship between size and restoration and how much it varies compared to my null hypothesis. Afterwards I will check the p-values to understand how likely it is that the relationship described if the null hypothesis of no relationship is true. If the test is more likely than the null hypothesis, I can infer that exists a statistically significant relationship between size and restoration time. If the test however is less likely than the null hypothesis, I can infer that there is no statisitcally significant relationship (Bevans 2022)

Results

Missing data

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.

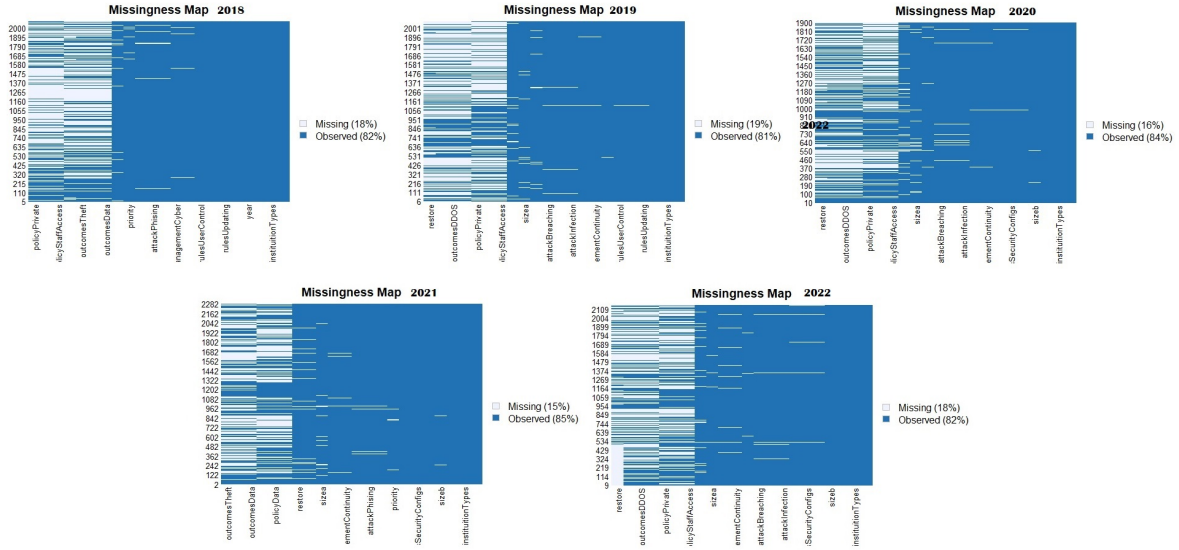

Figure 1: Missingness Map from 2018 to 2022

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. (White, Royston, and Wood 2010)

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 convergence was achieved after multiple trials with different numbers of iteration. ("Book_MI.knit" 2022)

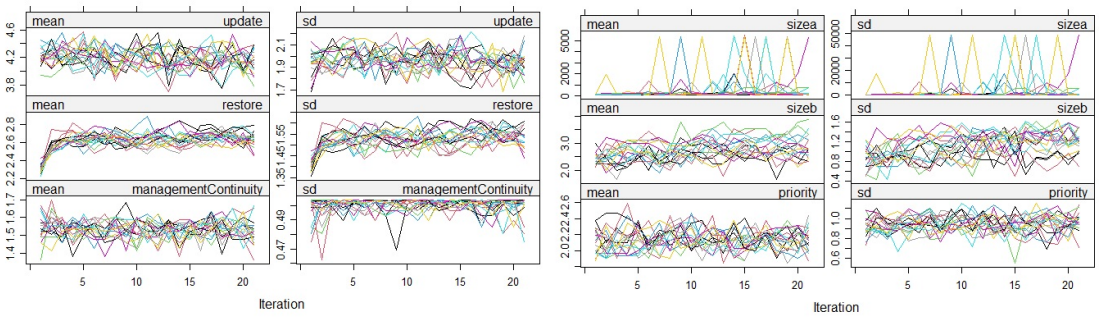

Figure 2: Healthy convergence plot

specify the size b and a imputations

The prediction matrix is a matrix which tells mice which variables can be used to predict missingness in the other variables. Mice by default uses the correlation between and the proportion of usable cases. For the prediction of the exact number of employee however prediction based only on the scale of size of the institution to avoid predicting values outside of the already known scale level of the institution when imputing the missing values.

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. ("Book_MI.knit" 2022)

Visual analyses

After dealing with the missing data and having a complete data set we can start our exploratory analysis. This analysis require us to first visualize the data to find any obvious patterns or groupings. Given the nature of the data a box plot will

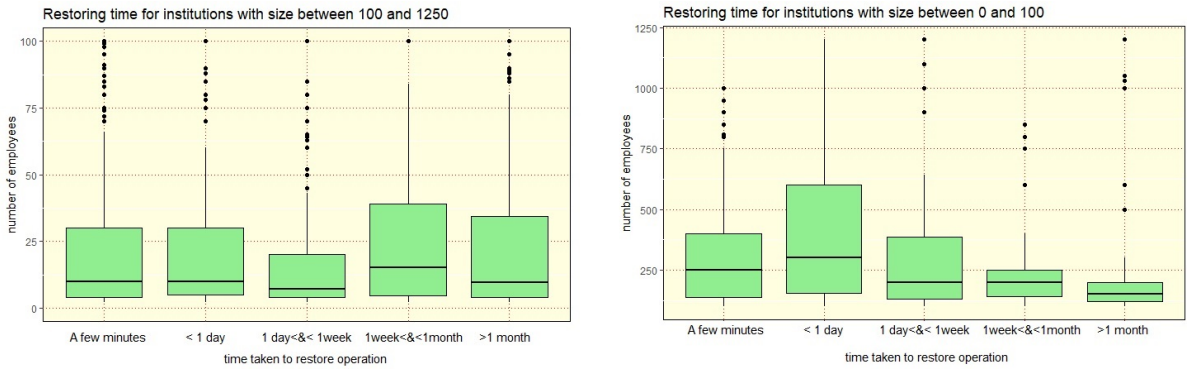

Figure 3: The box plot of the institution size compared to restoring time

hypothesis testing

To test my hypothesis I started with a simple mean comparison test between the size of the company and each of possible times it took to restore business operations.

Since the restoring time is recorded in multiple scales I cannot use a normal t-test, so I have chosen to use the Analysis of Variance (ANOVA) test to prove my hypothesis. Anova is a statistical test that compares the mean of multiple groups, in this case I have used one-way ANOVA since I am only comparing one one categorical independent variable with 5 levels that is the restoring time take and one quantitative dependent variable, the size of the organisation.

ANOVA output explains how much variation in the dependable variable can be explained by the independent variable, so how much does the time taken to restore affects the size the of the company.

TODO TABLE WITH ANOVA RESULTS

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, it is highly likely that the amount of cyber attacks is much higher since it is only possible to report the discovered cyber attacks.(Department For Digital 2020)

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. Another source of missing data is from the employees who participates in this survey and exercised their right to not answer some of the questions.

Furthermore the amount of missing data in each of the data sets is significant enough that if the imputed values were replaced with the real data the results could be considerably different because imputed data is not real data and does not account for any biased missing data contributing factor.

Lastly it would be possible to compensate for some bias related to the size of the institutions by implementing weighting to better represent the proportion of the smaller institutions.

Conclusion and recommendations

Bevans, Rebecca. 2022. "Choosing the Right Statistical Test: Types &Amp; Examples." *Scribbr*. https://www.scribbr.com/statistics/statistical-tests/.

"Book_MI.knit." 2022. *Home*. https://bookdown.org/mwheymans/bookmi/.

Department For Digital, Culture. 2020. "Cyber Security Breaches Survey, 2020." UK Data Service. https://doi.org/10.5255/UKDA-SN-8638-1.

Rubin, Donald B. 1975. "Biometrika 63 (3): 581–90." In *Inference and Missing Data*. Verlag nicht ermittelbar.

White, Ian R., Patrick Royston, and Angela M. Wood. 2010. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30 (4): 377–99. https://doi.org/10.1002/sim.4067.