Factors effecting and influencing cyber security precautionary measures

STA304 - Final Project

Uzair Mirza 1003657465

21/06/2021

Contents

1. Abstract

The goal of this study was to find the factors which influenced and had an effect on someone taking more precaution against cyber security threats. The data collected to conduct this was extracted from Statistics Canada, the dataframe was then cleaned and filtered used propensity score matching to lower the effect of bias. The main methods used to collect the results were; regression to see the effect of different factors along with propensity score matching to account for a treatment group, statistical tests were used to measure the significance of these factors. The result obtained translated to Incident our treatment being the most significant variable as expected followed by if someone shopped online having an increase in the precautionary measures one took.

2. Introduction

In the age of big data and digitization, cyber security and awareness about the measures people take for their privacy and protection about online presence is one of the greatest concerns in today's world.
During the time of COVID-19 most with most the activities going into remote operations and people working from home has led to an increase in the attacks and therefore the demand in awareness about cyber security and data privacy measures[**1**].

2.1 Goal

The goal of this study is to find the factors which effect and influence people and thus have resulted in people taking more precautions against cyber security threats.
The approach used for study being conducted is an observational approach. Observational study is when where we observe how a factor or factors are effected when compared to variations in other variables **without** trying to vary or influence the participants ie. the person conducting the study can not control the variables or choose the participants for the survey.[**2**]

2.2 Background, significance

Things related to technology have always made headlines and get people excited and happy but when it comes to cyber security incidents or data breaches they too get to make major headlines but have the opposite effect. All the way from recent data breaches, cyber security incidents which have had a tremendous effects globaly on both politics and economically[**3**] [**4**].
The importance of these concerns have led to the rise of the importance of cyber security and data privacy concerns, furthermore what is highly important is to observe what factors effect and give a rise in the precautions people take against such threats. Now due to COVID-19 and most of the activities are going into remote offering it is even more crucial to analyze and study these trends.

2.3 Hypothesis

We would expect and hope that people who have suffered from a related **incident or loss** would make an effort and now would be taking **more precautions** but the concern is that we cannot be sure and need to approach this problem using scientific methods. Thus to confirm this and to see the other factors which effect and lead to people taking more precautions we are conducting this study.

2.4 Data

The data collected to conduct this study is from a survey from Statistic Canada and is available to the public for free. The goal of the original survey was to measure the information about technology use, cyber security practices and online spending during the pandemic compared to before the pandemic[**5**].

2.5 Methods, terminology

To observe the relation between the variables we will be using **regression**, regression is used to see how changes in different factors effect the variable we are interested in, futhermore we will be using **propensity score matching** to filter out and match the data against our treatment group, which will be used to make our final model. Lastly to check the significance of these variables we will be using the result from the **z-test**.

2.6 Imediate Drawbacks

One of the biggest immediate drawbacks is when it comes to observational studies we don't have control over the factors effecting our results, this is most prominent when it comes to the variations between control vs treatment, however using propensity score might limit this bias but other issues such as confounding, and issues with validity and casuality still remain [**6**] [**7**].

3. Data

3.1 Data Extraction

As mentioned previously the data set is from Statistic Canada the reference about the data and all the different formats the data frame is available in can be found on their website. The direct link to download the dataframe along with the references can be found in the **appendix section 1.a**.
The direct download link will initiate the download of a zip file which includes the documentation about the variable codes and what they mean along with the codes for the answers for the questions and what they translate to.

3.2 Data Cleaning

Raw data directly from source had 3961 unique observations and 149 different variables of which 148 were the different properties and characteristics of each observation and 1 variable to uniquely identify each observation.
In raw form most of the variable were in code form so the first thing done was to use the documentation and filter our the variables in different categories. After filtering out and grouping the variables in different categories, translation of each value in each category was done using the available documentation. This was followed by different mathematical methods used to combine the related variables in each category into one unique category of interest. After this, filtareration of empty values(NA) was done to end up with individual categories of interests.
After having individual categories filtered out into the required formats these unique categories were merged together based on the original unique ID which was used to identify each observation in the data frame. The *flowchart of the exact steps* of the transformation and the different techniques used to get the data from source into the required form are described in **appendix section 1.b**.

3.3 Variables of intrest

Our main variable of interest is to see if one took more precautions to protect themselves, hence we have this category in form of *precaution* which has 2 levels *more* and *not more*. Now our goal is to see what were the significant factors which effected our quantity of interest.
Furthermore we also note that there is another driving factor which leads to someone taking more precaution, and this is whether or not they had an incident or loss related to data loss or a cyber threat. Hence we will be treating one of this variable as a treatment vs control group for our analysis.
However after cleaning the data set these are all the meaningful variables we are left with. Here first we have a table of all the variabales which are representative of the behaviours of the observations related to their internet and technology use;

Table 1: Important Variables

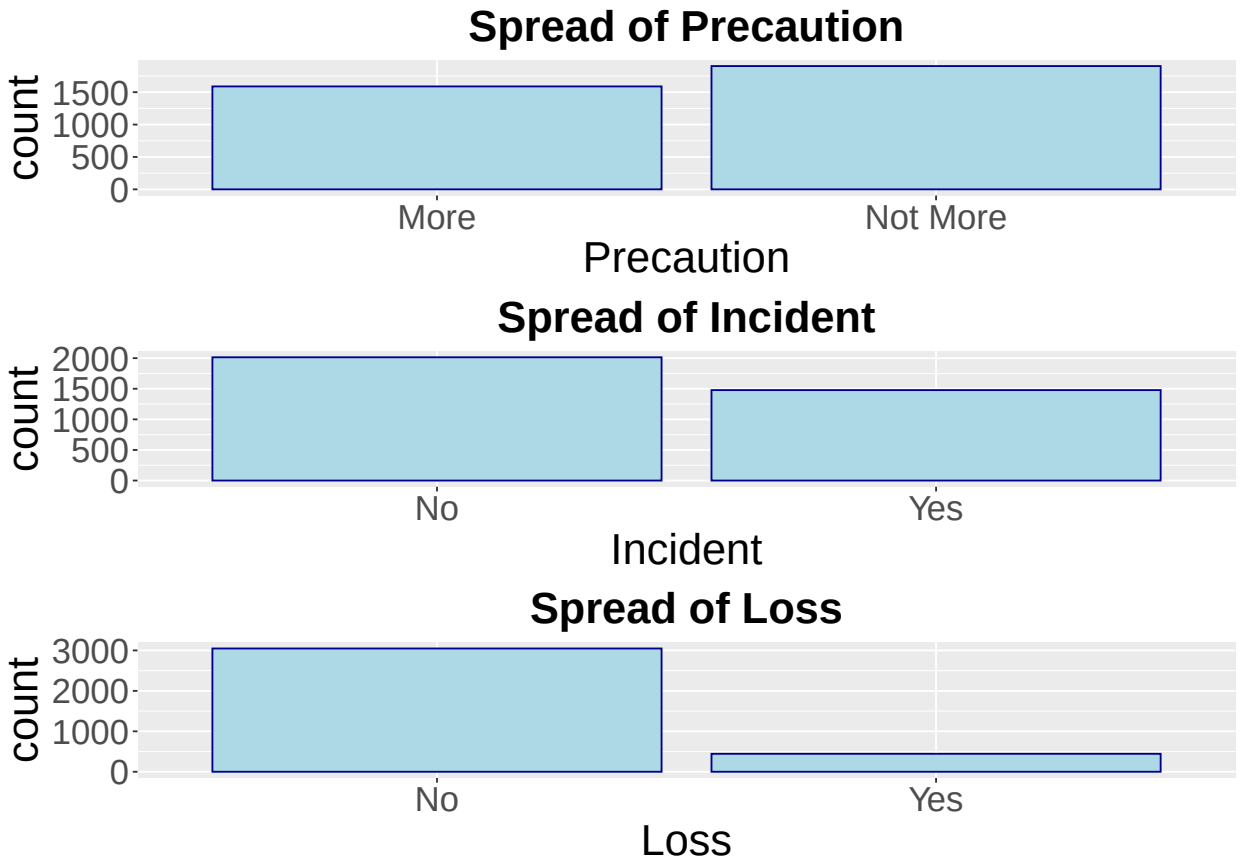
Variable	Meaning	Values
Online Shopping (binary)	Shopped online during pandemic?	Yes No
WFH (binary)	Worked from home during pandemic?	Yes No
Audio Streaming	Frequency of Audio Streaming use	Increase Same Decrease Not Applicable
Video Streaming	Frequency of Video Streaming use	Increase Same Decrease Not Applicable
Education Services	Frequency of Education Services use	Increase Same Decrease Not Applicable
Information Services	Frequency of Information Services use	Increase Same Decrease Not Applicable
Social media Use	Frequency of Social media use	Increase Same Decrease Not Applicable

And now we have all the variables related to the chracterstics of observations relating to the identification and features of the observation;

Table 2: Important Variables continued

Variable	Meaning	Values
Precaution (binary)	Took more precautions?	More Not More
Incident (binary)	Experienced an incident?	Yes No
Loss (binary)	Experienced a loss?	Yes No
Reported (binary)	Reported an incident?	Yes No
Household Size	The size of the individual's house-hold	1 2 3 4 4+
Age	The age group an individual belongs to	15 to 24 years old 25 to 34 years old 35 to 44 years old 45 to 54 years old 55 to 64 years old 65 to 74 years old 75 years and older
Sex (binary)	Biological Sex?	Male Female
Residence Area (binary)	Residential area classification	Rular Urban
Martial Status	The Martial status of an individual	Single/Never married Married Living common-law Widowed/Separated/Divorced
Education Level	The maximum level of education attained by the individual	Less than high school diploma or its equivalent High school diploma or a high school equivalency certificate College/CEGEP/other non-university certificate or diploma Trade certificate or diploma University certificate or diploma below the bachelor's level Bachelor's degree (e.g. B.A., B.Sc., LL.B.) University certificate, diploma, degree above the BA level

3.4 Plots and description



3.5 Plot Description

Bar Chart 1 Spread of Precaution:
Here we can see that the amount of people who took more precaution versus those who did not take extra precaution are fairly equally spread, with people who did not take more precaution having a slight higher frequency about 9.00% difference in both of their frequencies with people who did not take extra precaution having a 54.50% share versus 45.50% of that of the people who took extra precaution.

Bar Chart 2 Spread of Incident:
Here we can see that there is an increase in people who did not experience an incident versus the people who experienced an incident. There is about a 15.34% difference in both the categories with people who experienced an incident having a 42.33% share versus \$57.67% \$ of that of the people who did not experience an incident.

Bar Chart 3 Spread of Loss:
Here we can see that we have the most discrepancy between the the people who had a loss or did not have a loss. There is about a 75.60% difference between the two groups, with people who did not have an incident having a 88.32% share versus 11.68% share of that of people who had a loss.

3.6 Potential data issues

Here the biggest issue is in data cleaning when **combining related categories** into one category results in introduction of perhaps algorithm bias. Another issue with the data is the, **correlation between the variables** and how it will be effecting our estimates and we need to be careful about this.
Correlation refers to dependency between the variable, note having a correlation between your explanatory variables(factors) means that there is a relation between them however when observing the effects on our response variable we would ideally want that the explanatory variables be independent of each other. Example incident and loss have a correlation and thus we need to account for this when making our estimates.

4. Methods

As mentioned earlier we will be using regression to estimate the effect of how different quantities(explanatory variables/independent variables) interact with our variable of interest (response variable/dependent variable). So here our response variable is **precaution** which as mentioned in the previous table accounts for whether there was an increase in person's precaution against data-privacy and cyber security threats. All the other variables are the explanatory variables.

4.1 Logistic Regression

Now as precaution is a **binary variable** hence we will be doing **logistic regression** to account for this. Logistic regression is used when the response variable is binary, this is used over normal regression as the number of possible responses are between 2 categories(binary) compared to more than 2 categories or a range of values in normal regression model. The details about logistic regression are present in **appendix section 2.a**

4.2 Treatment, Control

The next thing we need to account for a possible **treatment group**. Treatment group in a study is a variable in our study which we assume is mainly effecting our response variable. Now for each observation in treatment group we need an observation from the **control group**. Now a control group is a group in our study which only differs from the treatment group by the treatment not being received. We will get back to matching the treatment with control after identifying the treatment variable in our study.
We identify that **loss and incident** as the 2 candidate of a possible treatment group, this is because research has shown that after an incident especially after a loss on average a person starts taking precaution, this is also attributed to the facts that studies have attributed that purpose of memory is stop one from making mistakes and hence taking more precaution comes into play after an incident or a loss.**[8]**
When comparing the two variables for the choice of treatment group we want to see the difference between the people who experienced a loss and check whether they took more precaution and compare that with the people who had incident against whether they took more precaution or not. Now to compare which group is more significant we do a statistical test to compare the effect of each treatment and see if the control is independent from treatment. To achieve this we do a Chi-Square Test of Independence, the result of running the test and comparing the results between incident and loss we identified that **incident** is a better treatment group for our study. Furthermore we also choose incident over loss as in our dataframe we have a higher frequency of people who experienced an incident over a loss, hence this would result in a larger dataset to make our model.

4.3 Matching

After identifying the treatment group we need to *match* each observation in treatment with a control. The reason for matching is because this is an observation study we cannot choose and select the observations for our control group. To get the closest observation for each of our treatment we will be doing **propensity score matching**. Propensity score matching is when we assign each observation in our data a score based around recieving treatment. In our case we made a logistic regression model for *incident* against the other variables to get a score for each observation. After assigning the score we match the each treatment with the closest observation with a similar score to control.
This results in a reduction of our original data set, with this new data frame being a 1 to 1 match for treatment and control. Now to check the quality of matching we go back to the Chi-Square Test of Independence to test the quality of matching, running the test we can see that the quality of matching is good for the study. The complete details about the reason, procedure of matching along with checking **quality of matching** can be found in **appendix section 2.b**.

4.4 Model selection

Now once we have the matched data we now go back to testing for our response variable. As mentioned before we will be making a logistic regression model for our response variable. Now our **choice for** the variables to choose for the **best model** is done using **Likelihood-ratio test**.
Likelihood-ratio test is used to test and compare if there is difference between 2 models which differ from one and another by 1 varible or 1 order in a variable, details about the Likelihood-ratio test and interpretation of the results can be found in **appendix section 2.c**.
The design of these models and the choice of these models is done by adding and removing different variables and comparing the results from the test to come up with the best model. Furthermore note that before making the model we need to remove the variables which are correlated to each other as they will be effecting our model.
Recall as mentioned before 2 variables are correlated when there is a relationship between them ideally when making the model we should aim that the variables making our model are independent of each other.

4.5 Variable significance

Now once we have the best model, to check for the significance of the variables in the current model we will be doing a z-test. The result from the z-test will tell us if the variable is significant or not based on the p-value we will be calculating our results. Now if the p-value for the variable < 0.05 we will consider that variable a significant variable, details about the calculation for the z-test from the regression model output can be found in **appendix section 2.d**.

A summary and flowchart of applying the approach can be found in **appendix section 2.e**.

4.6 Method drawbacks

The most immediate drawback is using propensity score for matching our treatment with control. Note when matching we are picking up observations on the basis of the score assigned to these observation however note that the control will have covariate values which will be further away from the treatment. Using this matched data to make our model results in an imbalance and hence leading our estimates being exposed to a bias[**9**].

5. Results

5.1 Best Model

To recall we started with a model with 17 variables of which 1 was our response variable, after that we started with a model consisting of all these variables and reduicng the number of varaibles on the basis of corelation and results from likelyhood ratio tests.

Based on the multiple models we find that the best model for predicting if someone took more precautions or not consists of the following variable;

Table 3: Variables for final model

Variable name
Video Streaming
Household Size
Age
Education Services
Incident
Information Services
Online Shopping

We can note that the variables which are in the final model are mostly the same variables we had identified in Table 1 as the representative of the behaviours and frecuency of internet and technology use. This is something not surprising as most of these interactions require the most data input and transfer. Furthermore another thing to note however is that frequency and use of Audio services is not a significant variable, this can be attributed to the fact that we had no variable to account for the free vs paid audio services. Having another level of interaction and having information about that would have possibly resulted in audio services being a significant variable aswell.

More details about *models along with the different variables* can be found in **appendix section 3.a**.

5.2 Estimates, Significant Variables, test results

In the table below we can find all the different categories of the variables along with the estimate and the significance(p-value) of the category.
We might notice that for each variable we are *missing a category*, this is because the estimate for this category is *adjusted in the regression intercept of the model*, which is β_0 in our regression equation for our model, furthermore the choice for this category is done on an alphabetical basis. Example; "No" comes before "Yes" hence category "No" for the variable is adjusted in the intercept for the model.

Table 4: Significant categories for variables and their estimates

Variable category	Estimate	p-value
Intercept	0.412675	0.43839
Video Streaming: Increase	-0.283354	0.27888
Video Streaming: Not Applicable	-0.302079	0.28289
Video Streaming: Same	-0.549102	0.03445*
Household Size: 2	0.170107	0.05522
Household Size: 3	0.151848	0.31211
HouseholdSize: 4	0.517397	0.00991*
HouseholdSize: 4+	0.119527	0.74877
Age: 25 to 34 years old	-0.514853	0.02018*
Age: 35 to 44 years old	-0.451976	0.03590*
Age: 45 to 54 years old	-0.207115	0.32568
Age: 55 to 64 years old	0.152017	0.47550
Age: 65 to 74 years old	0.213543	0.33823
Age: 75 years and older	-0.005908	0.98233
Education Services: Increase	-0.445458	0.20274
Education Services: Not Applicable	-0.490735	0.15129
Education Services: Same	-0.716208	0.03820*
Incident: Yes	0.554100	5.07×10^{-13}*
Information Services: Increase	-0.095966	0.82141
Information Services: Not Applicable	-0.212878	0.65396
Information Services: Same	-0.546570	0.19419
Online Shopping: Yes	0.545710	6.52×10^{-06}*

In the previous table values under the p-value which have * have a p-value less than 0.05, meaning they have a significant effect on th precaution a person takes and *thus are significant* in predicting for if somone took more precautions during this time or not.
Lastly note that the **Estimate** in our previous table translates to the effect that catagoery has on the log-odds of someone taking more precaution or not. This means that increase in estimate is directly propotional to the increase in the precaution a person would take. The reason for this can be found in **appendix section 3.b**.

The highlights of the results we can see that as expected and identified having an Incident is significant and also increases the precautions one takes. Likewise is the case with online shopping, consuming video streaming and being in an age group of mid 20s to low 40s, however we can see that having a house size of 4 is also significant BUT has the opposite effect.

6. Drawbacks

No method is foolproof or 100% perfect this is also the case in our study, let us look at a summary and the highlights of the issues.
As mentioned previously the first drawback of our study is that this is an observational study and hence we don't have a control on our observations, this bleeds into the problem which arrises with matching as that might be adding much more bias to our model.
The other drawback is that when combining variables of similar category we might be introducing our own bias with the algorithm we choose to combining them, hence introducing more sources of bias. The other major drawback in our study is we are not acccounting for a very strong dependant of our response variable which is during the time of the pandemic and when people started working remotely some of them had standard precautionary protocols from work which required them to take certain measures and hence increasing the precautions they took.
Lastly we are not checking or accounting for how much the survey data used to conduct this study accounts for the true representation of the population.

7. Conclusion

In conclusion we started our observation study with first choosing our data frame and cleaning and converting our dataframe. After this we identified the presence of a treatment group and did propensity score matching to get the control for each treatment, this was followed by checking the quality of our matching.
Once we have a matched dataset we make different models and compare them to check which model is the best model and hence telling us which variables are the best predictors for someone taking more precautions or not.
Lastly to check the significance for the variables and their categories we did a z-test to see which variable and their categories are the most significant ones.
The result we found was that yes having an incident is a strong predictor along with if someone is online shopping or not, which makes sense as there is an imediate risk associated with it.
Furthermore we also noted some drawbacks of observational studies, along with that we also noted the drawbacks from using propensity score matching along with our alogirtm biases when combining categories, lastly we also should highlight how not accounting if someone had directions from work can have a strong effect on our study.
Looking further, having accounted for if not all then most of the drawbacks one can try doing a poststratification on a similar dataset or divide the datset in a 90:10 split to train and to check the quality of the model and after that account for the most significant variables.

8. Biblography

1. Cyber Risks: An Increased Threat During COVID-19. (n.d.). Insurance Bureau of Canada. Retrieved June 21, 2021, from http://www.ibc.ca/ns/business/risk-management/cyber-risk/an-increased-threat-during-covid-19
2. Observational vs. experimental studies. (n.d.). Institute of Work & Health. Retrieved June 21, 2021, from https://www.iwh.on.ca/what-researchers-mean-by/observational-vs-experimental-studies#:~:text=Observational%20studies%20are%20ones%20where,two%20types%20of%20observational%20studies.
3. 5 Tensions Between Cybersecurity and Other Public Policy Concerns. (n.d.). Sciences Engineering Medicine. Retrieved June 21, 2021, from https://www.nap.edu/read/18749/chapter/7#105
4. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. (n.d.). The Guardian. Retrieved June 21, 2021, from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election
5. Canadian Perspectives Survey Series 5: Technology Use and Cyber Security During the Pandemic Public Use Microdata File. (n.d.). Statistics Canada. Retrieved June 21, 2021, from https://www150.statcan.gc.ca/n1/pub/45-25-0010/452500102021001-eng.htm
6. Observational Studies: Uses and Limitations. (n.d.). SpringerLink. Retrieved June 21, 2021, from https://link.springer.com/chapter/10.1007/978-3-319-99124-5_31#:~:text=Observational%20studies%20are%20a%20
7. Observational Studies: Uses and Limitations. (n.d.-b). SpringerLink. Retrieved June 21, 2021, from https://link.springer.com/chapter/10.1007/978-3-319-99124-5_31#:~:text=Observational%20studies%20are%20a%20
8. 12 Rules for Life An Antidote To Chaos. (n.d.). Random House Canada.
9. Why you shouldn't use propensity score matching. (n.d.). The Stats Geek. Retrieved June 21, 2021, from https://thestatsgeek.com/2016/09/07/why-you-shouldnt-use-propensity-score-matching/

9. Appendix

1. Data

1a. Data extraction

The reference and source along with the direct download link to get the zip file containing the data in a csv format along with the documentation can be found in the table below

Table 5: Data source and download link

Refrence	URL link
Statistic Canada, Refrence about the Data source	https://www150.statcan.gc.ca/n1/pub/45-25-0010/452500102021001-eng.htm
Direct download, dataframe and documentation	https://www150.statcan.gc.ca/n1/pub/45-25-0010/2021001/CSV-eng.zip

1b. Data Cleaning

Firstly using the documentation the related categories were combined and separated. In general there were 3 different types of techniques which were used to group these categories together;
Technique 1 was for categories like Sex, Age-group, Education-level, and other similar categories which just required the translation from codes to what the codes actually mean.
Technique 2 was for categories which were spread across multiple categories and had binary options as the quantity of interest(Yes, No or NA) in the original data set and required to be filtered and combined into 1 category with binary options(Yes, No).
Technique 3 was for again for categories which were spread across multiple categories but had multiple quantities of interest(Less, More, Same or NA) and had to filtered and merged into one category with a binary outcome(More, Not More).
Let us now look at each example for each one of this category;

Technique 1;
Let us look at one of the variable of interest which covers all the multiple cases which are dealt with this technique. This variable is the frequency of social media use, first we identify the question that deals with this variable in this case it is "CPD_05A" in the original dataframe after that we first selected this column along with the unique ID to identify each observation ie. "PUMFID". Now that we have a unique identifier against the category of interest we now translate the numeric codes(1,2,3,4,9) into meaningful categories(Increase, Same, Decrease, Not Applicable, Skip) and saved it into the dataframe. After that we stored these new translated categories along with the original unique identifier into a new dataframe containing just the 2 meaningful variables which are the unique observation identifier along with the translated meaningful categories.

Technique 2;
Now for this technique let us look at the variable which tells us if someone experienced some incident related to cyber security or data privacy concerns or not. Again using the documentation we first identify the related columns, we select those columns along with the unique observation identifier as before. In this case we we extracted the first 13 columns as they translated to all the different variables related to different forms of incident. Here the possible valuable answers were(1,2,9) which when translated would be translated to 1 being a Yes 2 being a No and 9 being a skip.
Now our goal is to find is anyone had any form of an incident and record it, to account for this the first thing we did was change the original code 2(No incident) to a 0 and then we took a sum across the row and recorded it.
Now if someone skipped for all the questions related to the incident the sum would add up to $117(9 \times categories = 9 \times 13 = 117)$ so if the sum of the row is 117 we *filter these values out.*
Similarly to account for if someone did not face any incident their sum would come to 0, this is because recall we first we changed the value 2 to a 0 and now the row sum would be $0(0 \times categories = 0 \times 13 = 0)$ hence we assign that row to *Not having an incident.*
Now to assign the value for if someone had an incident similarly their row sum would be between greater than 0 and less that 117, infact the bounds for this would come up to $\Sigma_1^i 1, i \in [1, 13]$ which results the sum being in between 1 to 13. Hence this sum would be assigned a value of *had an incident* in our dataframe.
Lastly again just like before we are only going to keep and record this new variable we created along with the unique observation identifier.

Technique 3;
Lastly for this technique let us look at the category which will be measuring if someone took more or not more Precaution related to cyber security threats or data privacy concerns. Here agaun just like in the other techniques we will be filtering out the categories related to this variables in total there were 10 variables along with the unique observation identifier.
Now here in the raw form the possible entries are(1,2,3,4,9) which translate to 1 meaning more, 2 meaning about the same, 3 meaning less, 4 meaning not applicable and 9 meaning skip.
Now our goal is to identify if someone took extra precautionary measures in any form and also identidy and filter out the observations who skipped the entire group of questions in the survey.
The first thing done was changing the 1(More) to a 0 and 4(Not applicable) to 5, after that we first filtered out the entries who skipped for every category this was done by **filtering out** the observations which had a row product of 3486784401 this is because $9^{categories} = 9^{10} = 3486784401$. After that we **filtered out** the people the people who were **not applicable** for anyone of these categories their row product would be 9765625 this is because $5^{categories} = 5^{10} = 9765625$.
Lastly now we group our remaining dataset in **more and not more** on the basis of the row product of 0 translates to More and row product not equal to 0 translates to Not more.
Lastly again just like before we are only going to keep and record this new variable we created along with the unique observation identifier.

Lastly we now **merge** all these different categories we created using either one of the 3 techniques together on the basis of the unique observation identifier. This will filter out and give us a clean dataset with all the required categories.

Now our cleaned data sets looks like this;

Data Glimpse

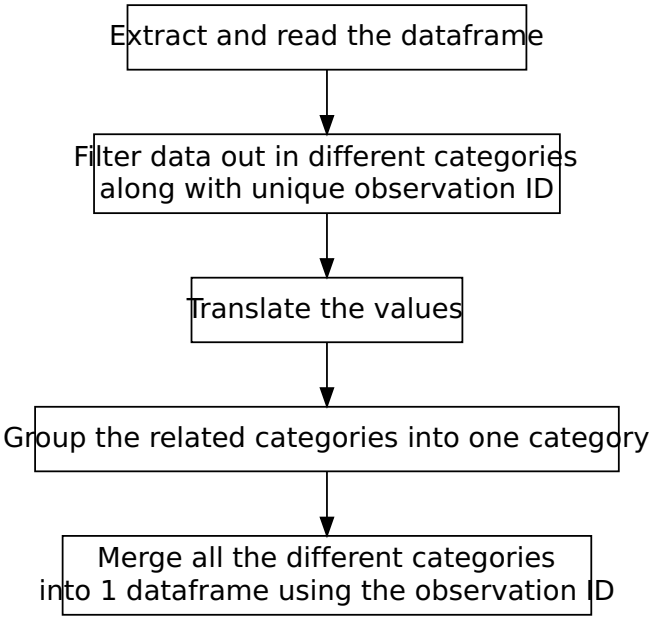
```
## # A tibble: 6 x 4
##   Audio_Streaming Video_Streaming HouseholdSize Age
##   <chr>           <chr>           <chr>         <chr>
## 1 Increase        Increase        3             35 to 44 years old
## 2 Not Applicable  Same            2             65 to 74 years old
## 3 Same            Same            2             65 to 74 years old
## 4 Not Applicable  Same            1             55 to 64 years old
## 5 Same            Same            3             65 to 74 years old
## 6 Same            Same            2             45 to 54 years old
```

```
## # A tibble: 6 x 6
##   Martial.Status  Education.Level   Sex   Residence.Area Precaution Edu_Services
##   <chr>           <chr>             <chr> <chr>          <chr>      <chr>
## 1 Living common-~ Trade certificat~ Male  Urban          More       Not Applica~
## 2 Married         University certi~ Fema~ Urban          Not More   Not Applica~
## 3 Married         University certi~ Male  Urban          Not More   Not Applica~
## 4 Single/Never m~ College/CEGEP/ot~ Fema~ Rural          More       Not Applica~
## 5 Married         Bachelor's degre~ Male  Rural          More       Same
## 6 Married         College/CEGEP/ot~ Male  Rural          Not More   Same
```

```
## # A tibble: 6 x 5
##   Reported Incident Inform_Services Loss  OnlineShopping
##   <chr>    <chr>    <chr>           <chr> <chr>
## 1 Yes      Yes      Same            No    Yes
## 2 Yes      Yes      Same            Yes   Yes
## 3 No       No       Same            No    Yes
## 4 No       No       Increase        No    Yes
## 5 No       Yes      Increase        No    Yes
## 6 Yes      Yes      Same            No    Yes
```

```
## # A tibble: 6 x 2
##   Social_Media_Use WFH
##   <chr>            <chr>
## 1 Same             No
## 2 Same             Yes
## 3 Not Applicable   Yes
## 4 Increase         No
## 5 Decrease         Yes
## 6 Same             Yes
```

In summary this is a flowchart of all the steps being taken to clean the data.



2. Methods

2a. Logistic Regression

The equation for the logistic regression model is given by;

$$\log(\frac{\hat{p}_{h_0}}{1-\hat{p}_{h_0}}) = \beta_0 + \Sigma_{i=1}^{k}\beta_i x_i \quad k \subset [1, n]$$

Here is a breakdown of this equation;
$\log(\frac{\hat{p}_{h_0}}{1-\hat{p}_{h_0}})$ will give us the logs odds for our response variable.
$\hat{p}_{h_0}$ is the probability of our binary response variable(**precaution**).
β_0 is the intercept of our regression model.
β_i is the Estimated change in the response variable's odds against the i'th variable which is fixed for the model(**explanatory variable coefficient**).
x_i is the varying i'th variable's value for each individual observation.
n is the number of of different categories which we use to make the model.

2b. Matching, Propensity score, quality of matching

To assign a score for each observation we first made a logistic regression model similar to that in appendix 2.a but now the treatment variable, which is incident is the response variable. Now after making the model we get the fitted values(score) for each variable by getting the probability of each observation being in the treatment group.
From the model which generates log odds to probability we get there using the following steps;
STEP 1:

$$\log(\frac{\hat{p}_{h_0}}{1-\hat{p}_{h_0}}) = \beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i \quad \implies \frac{\hat{p}_{h_0}}{1-\hat{p}_{h_0}} = \exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i)$$

STEP 2:

$$\hat{p}_{h_0} = (1-\hat{p}_{h_0})(\exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i)) \quad \implies \hat{p}_{h_0} = (\exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i)) - \hat{p}_{h_0}(\exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i))$$

STEP 3:

$$\hat{p}_{h_0} + \hat{p}_{h_0}(\exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i)) = \exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i) \quad \implies \hat{p}_{h_0}(1 + (\exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i)) = \exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i)$$

STEP 4:

$$\hat{p}_{h_0}(1 + (\exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i)) = \exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i) \quad \implies \hat{p}_{h_0} = \frac{\exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i)}{(1 + (\exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i))}$$

So score we have is:

$$\hat{p}_{h_0} = \frac{\exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i)}{(1 + (\exp(\beta_{0j} + \Sigma_{i=1}^{k}\beta_i x_i))}$$

As mentioned before this generated probability ie. $\hat{p}_{h_0}$ is the score being assigned to each value and we then use the closest/similar score around the treatment to match the control.

Now to check the quality of matching we will be doing a Chi-square independence test. We use a Chi-square independence test to test if there is a relationship between 2 categorical variables. In our case it is Incident vs No Incident and see if there is a relationship between them when it comes to the level of precaution. For this test the H_0 which is the Null Hypothesis is that there is no difference between people who had an incident(treatment) vs no-incident(control) and Alternative Hypothesis H_a is that there is difference between the 2 groups. Running the test we get a result of p-value of 2.31×10^{-13}. Now as the p-value $<$ 0.05 this translates to rejecting the H_0 meaning that there is a significant difference between the treatment and control and hence meaning the quality of our matching is sufficient.

2c. Likelihood-ratio test and how to apply in our case

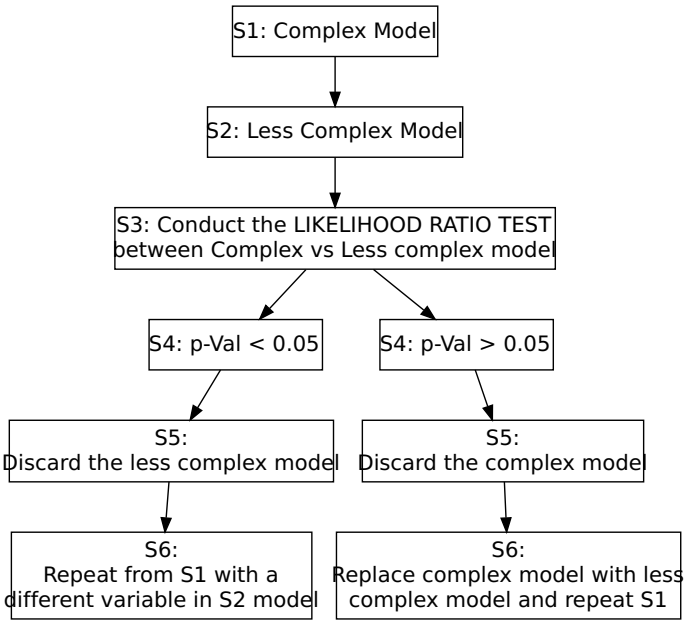
When comparing 2 different models which differ from the each other by 1 variable or 1 order we use a likelihood-ratio test, this test assesses the goodness of the fits based on the ratio of their likelihoods, we test if the ratio is different from one, or if the natural logarithm of the ratio is different from zero. Note we are essentially comparing a more complex model against a less complex model which differs from the complex model by 1 variable or by 1 order in a particular estimate.
For this test the H_0 is that there is a difference between the 2 models ie. the smaller less complex model provides an almost equal fit as the more complex model with more variables and H_a is that there is a difference between the 2 model. So to apply this in our case we start with a complex model with all the variables and then keep dropping variables and keep running the test to see which model along with the variables is the most significant variables.

2d. Variable significance, z-test calculation

To test for the significance we will be using a z-test and getting the p-value for generated z-score, now this z-score is generated from the output of our our regression model.
For each variable our regression model will generate an **Estimate**, followed by the **Standard error**, **Z-score** and the **p-value** for the z-score. Note here the Estimate is the effect of the variable on the log-odds for *precaution* the standard error tells us how wrong our regression estimate is, smaller the standard error the better it. Now the generated Z-score is the ratio of Estimate against the standard error. The method of the working of the test is similar to that of Chi-square test of independence between different proportions and hence use the same null and alternative hypothesis from **appendix 2b**, and similarly having a p-value less than 0.05 would mean that our variable is significant as there is difference when comparing different proportions.

2e. Summary of Methods and how to apply them

Below is the flowchart for steps to apply to get to the best model, we keep repeating this cycle until we have tried all the possible variable combinations. Once we have the best model we go on to the significance test for the variables in the model to identify the most significant variable effecting the precaution level.



3. Results

3.a Model selection steps taken

We started with the initial model with all the variables except loss in our initial model as the most complex model. After that we removed the corelated variables, we found *incident* to be corelated with our treatment so we remoeved that. After that we made in total about 14 different models to get to the final model. We followed the flowchart presented in *Section 2.e* to get our results. In general we first got rid of the corelated variables which were again as mentioned incident and also reportted which is if someone reported the incident or not.
After getting rid of the corelated variables we kept making different models with different variables and tested their results using the likelyhood ratio test. The results resulted is us getting rid of, Work from home, Social Media Use, Residence Area, Martial Status and Education level as having models with all these variables had little to no effect on the significance and the effectiveness of the predicting if someone took more precaution or not.

3.b Odds relations with our hypothesised probability

$$\hat{p}_{h_0} = \frac{\exp(odds)}{(1 + exp(odds))} \quad \implies odds \propto \hat{p}_{h_0}$$

Here $\hat{p}_{h_0}$ is the probability of someone taking more precautions. So we can see odds which is sum of all the coeficients in the model is dorectly propotional to the increase in the probability of precaution one will be taking.