

# Exploring Past Year Identity Theft with the 2016 Identity Theft Supplement of the National Crime Victimization Survey

Erika Harrell

12/21/2020

## Executive Summary

Data analysis was conducted to see if certain predictor variables were associated with past year identity theft. The data used in this analysis came from the 2016 Identity Theft Supplement (ITS) to the National Crime Victimization Survey (NCVS). The results of this analysis found that identity theft was dependent on the majority of the predictors used (age, race/Hispanic origin, annual household income, being a victim of identity theft prior to the past year, having personal information exposed due to a data breach, and using preventative behaviors). Gender was found to be independent of past year identity theft ( $p>0.5$ ).

```
#adding libraries
```

```
library(dplyr)
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
## filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
## intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
library(ggthemes)
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
```

```
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##      recode

library(ROCR)
library(Metrics)

##
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':
##
##      precision, recall

library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following object is masked from 'package:Metrics':
##
##      auc

## The following objects are masked from 'package:stats':
##
##      cov, smooth, var
```

## Loading Data

Data were downloaded from the 2016 ITS study page from the National Archives of Criminal Justice Data to a subfolder entitled “data” in the project on R-Studio, unzipped, loaded into R-Studio and renamed with a shorter name. Due to the size of the datafile, the memory limit had to be increased prior to loading the data in R-Studio.

```
## [1] 20000
```

## Data Wrangling

There were 125,165 total persons in the 2016 ITS. Only completed telephone and personal interviews were used in the analysis which left 96,130 interviews or observations in the dataset.

```
##                               Number
## (1) Personal interview      57119
## (2) Telephone interview    39011
## (5) ITS Noninterview       29035
## Total                      125165

##                               Number
## (1) Personal interview      57119
## (2) Telephone interview    39011
## (5) ITS Noninterview        0
## Total                      96130
```

Created variables that would be used in analysis from the larger ITS dataset.

Individual variables created in the previous step were combined into a smaller dataset and larger dataset was removed.

Summary of the created dataset.

```
## 'data.frame': 96130 obs. of 22 variables:
## $ sex : Factor w/ 3 levels "(1) Male","(2) Female",...: 2 1 1 1 1 2 1 1 2 2 ...
## $ race : Factor w/ 21 levels "(01) White only",...: 1 2 6 1 1 1 1 1 1 1 ...
## $ hispanic : Factor w/ 3 levels "(1) Yes","(2) No",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ income : Factor w/ 14 levels "(01) Less than $5,000",...: 14 14 14 14 4 13 13 13 13 1 ...
## $ age : num 46 50 22 78 50 30 29 62 60 74 ...
## $ pastyearbankacct : Factor w/ 5 levels "(01) Yes","(02) No",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ existing_bank : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ currentccacct : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 2 1 1 1 1 2 1 1 ...
## $ pastyearccacct : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 2 NA NA NA NA 2 NA NA ...
## $ existing_credit_card: Factor w/ 5 levels "(01) Yes","(02) No",...: NA NA NA 2 2 2 2 NA 2 2 ...
## $ other_existing_accts: Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ open_new_acct : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ personal_info : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ OUTSIDE_PAST_YEAR : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 1 2 2 1 2 2 2 2 ...
## $ CHCKD_CR_PAST_YR : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 1 2 2 1 1 2 2 1 ...
## $ CHNG_PASSWORDS : Factor w/ 5 levels "(01) Yes","(02) No",...: 1 1 2 2 2 1 1 2 2 2 ...
## $ PURCHASE_IDTHFT_INS : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ SHRED_DOCS : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 1 1 1 1 1 1 1 2 ...
## $ VERIFY_CHARGES : Factor w/ 5 levels "(01) Yes","(02) No",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ PROTECT_COMPUTER : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ PURCHASE_IDTHFT_PROT: Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ notify_breach : Factor w/ 5 levels "(01) Yes","(02) No",...: 2 2 2 2 2 2 2 2 2 2 ...
```

```
## sex race hispanic
## (1) Male :44908 (01) White only :79770 (1) Yes :12131
## (2) Female :51222 (02) Black only :10051 (2) No :83999
## (8) Residue: 0 (04) Asian only : 4160 (8) Residue: 0
## (03) Am Ind/AK native only: 656
## (07) White-Amer Ind : 530
## (06) White-Black : 304
## (Other) : 659
## income age pastyearbankacct
## (14) $75,000 and over :33662 Min. :16.00 (01) Yes :85793
## (13) $50,000 to $74,999:17342 1st Qu.:34.00 (02) No :10337
## (12) $40,000 to $49,999: 9330 Median :50.00 (08) Residue : 0
## (10) $30,000 to $34,999: 5811 Mean :49.26 (98) Refused : 0
## (11) $35,000 to $39,999: 5316 3rd Qu.:63.00 (99) Don't know: 0
## (08) $20,000 to $24,999: 5137 Max. :90.00
## (Other) :19532
## existing_bank currentccacct pastyearccacct
## (01) Yes : 4665 (01) Yes :69446 (01) Yes : 1003
## (02) No :81128 (02) No :26670 (02) No :25681
## (08) Residue : 0 (08) Residue : 0 (08) Residue : 0
## (98) Refused : 0 (98) Refused : 10 (98) Refused : 0
## (99) Don't know: 0 (99) Don't know: 4 (99) Don't know: 0
## NA's :10337 NA's :69446
```

```

##
##      existing_credit_card      other_existing_accts      open_new_acct
## (01) Yes      : 5460      (01) Yes      : 815      (01) Yes      : 589
## (02) No      :64989      (02) No      :95315      (02) No      :95541
## (08) Residue : 0      (08) Residue : 0      (08) Residue : 0
## (98) Refused : 0      (98) Refused : 0      (98) Refused : 0
## (99) Don't know: 0      (99) Don't know: 0      (99) Don't know: 0
## NA's      :25681
##
##      personal_info      OUTSIDE_PAST_YEAR      CHCKD_CR_PAST_YR
## (01) Yes      : 473      (01) Yes      :12267      (01) Yes      :44385
## (02) No      :95657      (02) No      :83692      (02) No      :51344
## (08) Residue : 0      (08) Residue : 48      (08) Residue : 80
## (98) Refused : 0      (98) Refused : 39      (98) Refused : 158
## (99) Don't know: 0      (99) Don't know: 84      (99) Don't know: 163
##
##
##      CHNG_PASSWORDS      PURCHASE_IDTHFT_INS      SHRED_DOCS
## (01) Yes      :36861      (01) Yes      :12149      (01) Yes      :67772
## (02) No      :58670      (02) No      :83440      (02) No      :27942
## (08) Residue : 88      (08) Residue : 88      (08) Residue : 92
## (98) Refused : 250      (98) Refused : 193      (98) Refused : 196
## (99) Don't know: 261      (99) Don't know: 260      (99) Don't know: 128
##
##
##      VERIFY_CHARGES      PROTECT_COMPUTER      PURCHASE_IDTHFT_PROT
## (01) Yes      :75419      (01) Yes      :16447      (01) Yes      : 4881
## (02) No      :20344      (02) No      :79001      (02) No      :90756
## (08) Residue : 94      (08) Residue : 100      (08) Residue : 100
## (98) Refused : 180      (98) Refused : 209      (98) Refused : 217
## (99) Don't know: 93      (99) Don't know: 373      (99) Don't know: 176
##
##
##      notify_breach
## (01) Yes      :11037
## (02) No      :84652
## (08) Residue : 102
## (98) Refused : 179
## (99) Don't know: 160
##
##

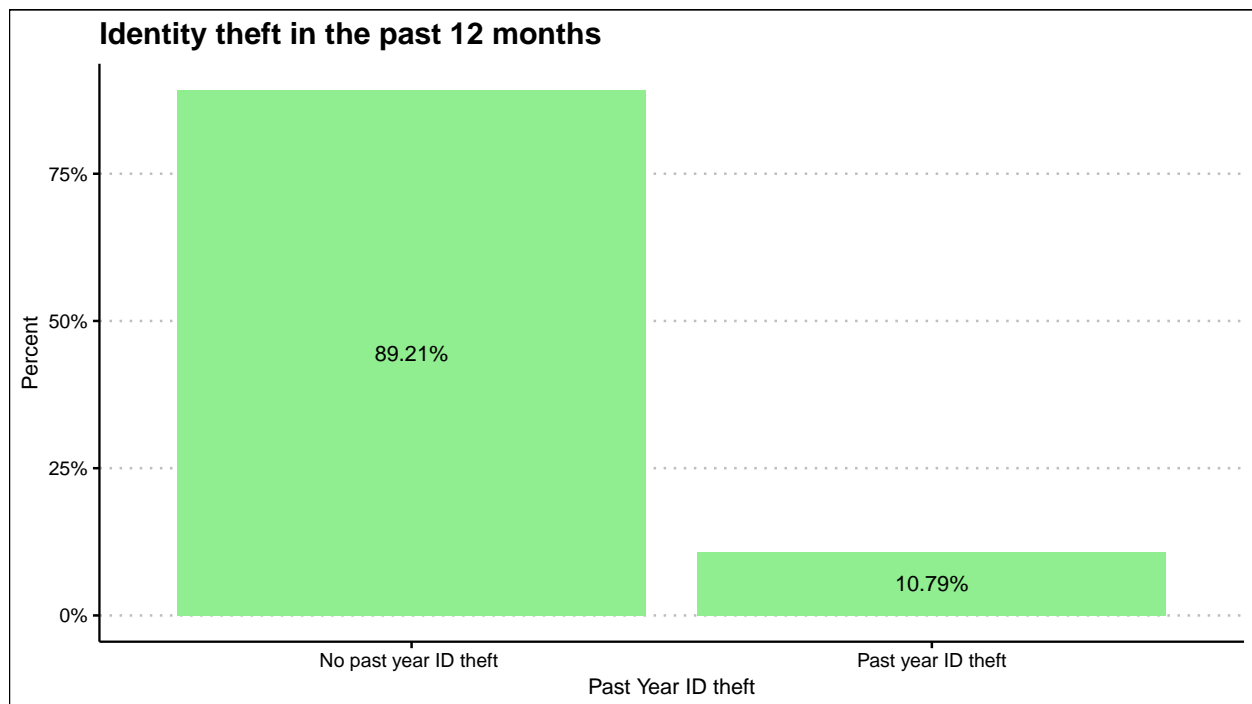
```

Recoded variables, collapsing categories and computing necessary categories and variables for analysis.

## Exploratory analysis

### Past year identity theft

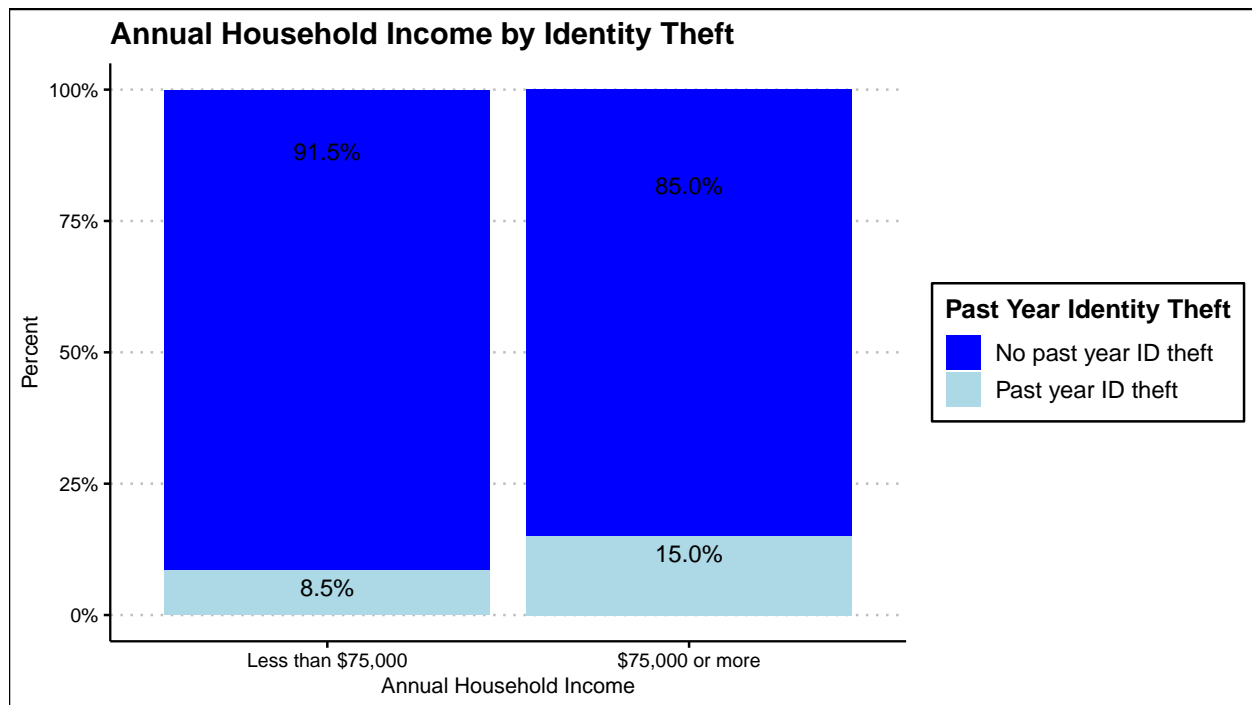
About 11% of the sample reported at least one type of identity theft (misuse of an existing account, misuse of personal information to open new account, or misuse of personal information for other fraudulent purposes) in the past year while 89% of the sample reported no identity theft.



### Annual household income

About two thirds of the sample were in households with annual incomes of less than \$75,000 (65%) while the remainder (35%) were in households with annual incomes of at least \$75,000. Within each income category the majority of respondents did not report experiencing identity theft in the past year. However, 15% of persons in households with incomes of \$75,000 or more reported past year identity theft, compared to 9% of those in other households.

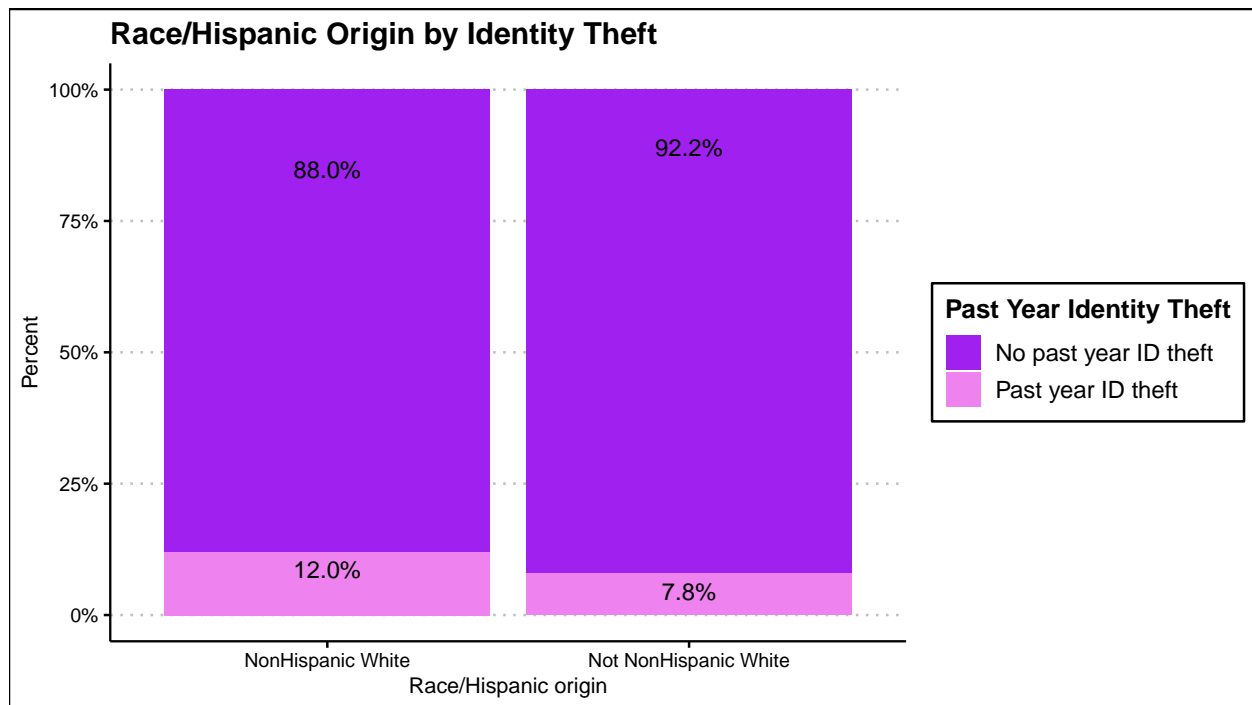
##	Count	Percentage
## Total	96130	100
## Less than \$75,000	62468	65
## \$75,000 or more	33662	35



## Race/Hispanic origin

Seventy one percent of respondents were NonHispanic White while 29% were belonged to other race/Hispanic origin groups. Twelve percent of nonHispanic Whites reported past year identity theft compared to 8% of other persons reported identity theft in the past year.

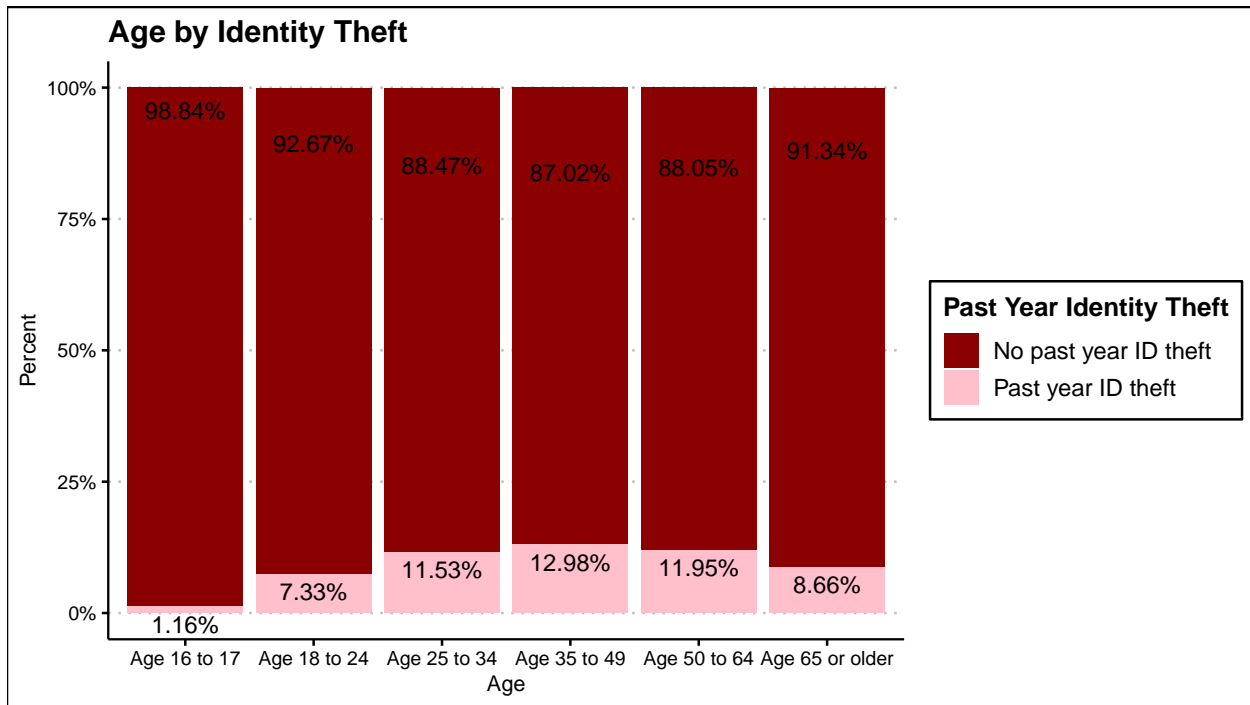
##	Count	Percentage
## Total	96130	100
## NonHispanic White	68265	71
## Not NonHispanic White	27865	29



## Age

Twenty eight percent of the sample was age 50 to 64 while nearly one in four (24%) were age 35 to 49. 23% of the sample was age 65 or older. The remainder of the sample was under the age of 35. Thirteen percent of persons age 35 to 49 reported experiencing past year identity theft, compared to 7% of persons age 18 to 34 and 9% of persons age 65 or older.

##	Count	Percentage
## Total	96130	100
## Age 16 to 17	1986	2
## Age 18 to 24	7826	8
## Age 25 to 34	14740	15
## Age 35 to 49	23147	24
## Age 50 to 64	26621	28
## Age 65 or older	21810	23

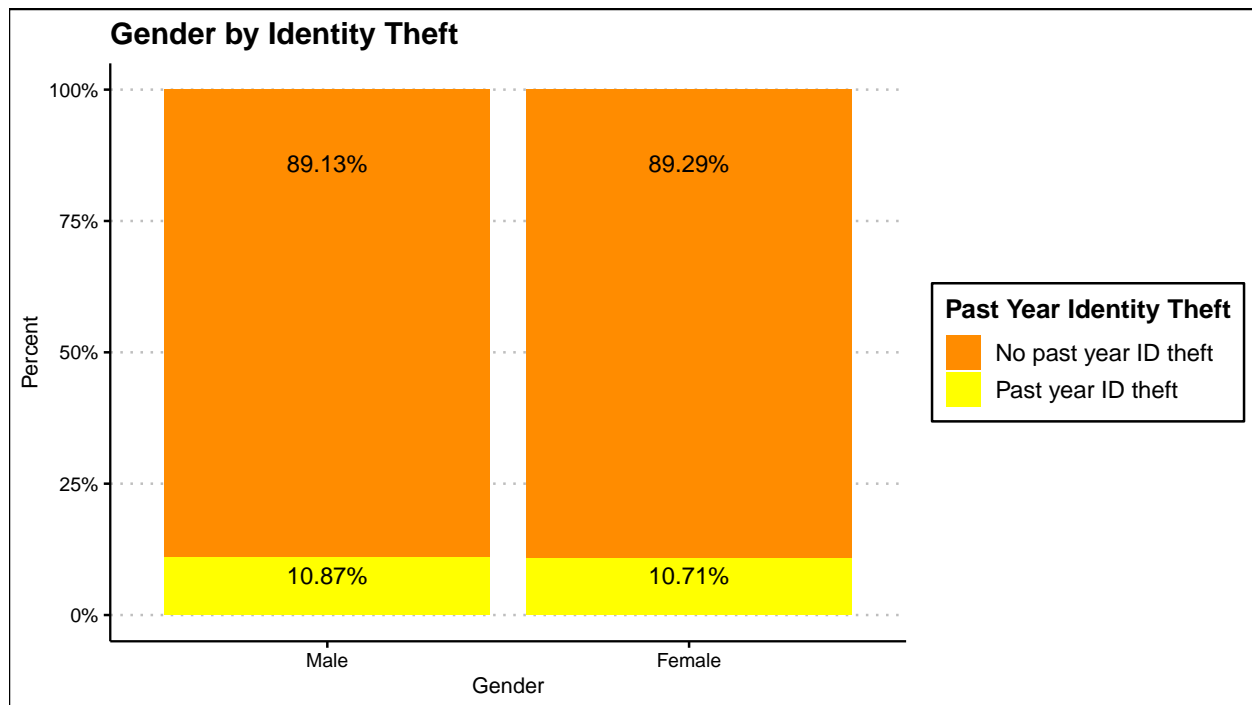


## Gender

More than half of the sample (53%) was female while the remainder (47%) was male. Past year identity theft was experienced by 11% of males and a similar percentage of females.

##	Count	Percentage
## Total	96130	100
## Male	44908	47
## Female	51222	53
## Unknown	0	0

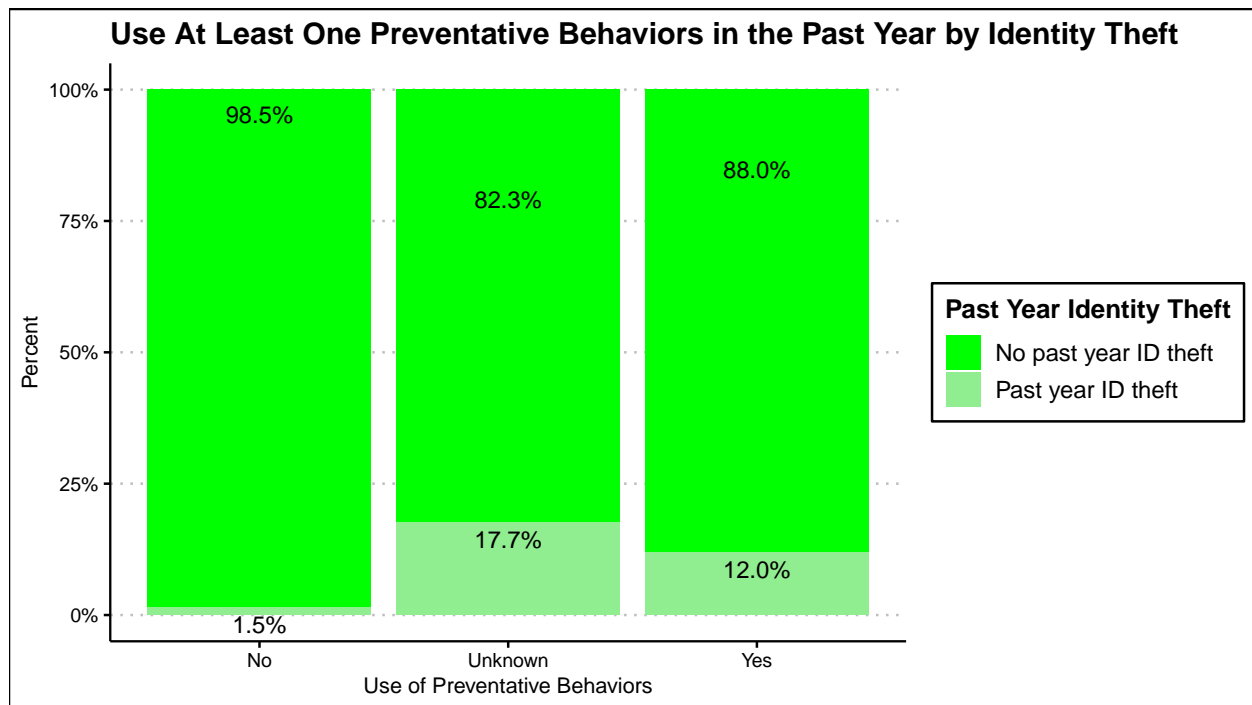




## Preventative behaviors

Nearly nine out of ten persons in the sample (88%) used at least one of the preventative behaviors measured (checked credit report, changed password on financial accounts, had credit monitoring services or identity theft insurance, shredded or destroyed documents containing personal information, checked bank or credit card statements for unfamiliar charges, purchased identity theft protection) in the past 12 months. Eighteen percent of respondents who did not know if they had used a preventative behavior in the past 12 months reported past year identity theft compared to 12% of those who had used at least 1 preventative behavior in the past 12 months. Surprisingly, only 2% of those who did not participate in any preventative behaviors in the past 12 months reported past year identity theft.

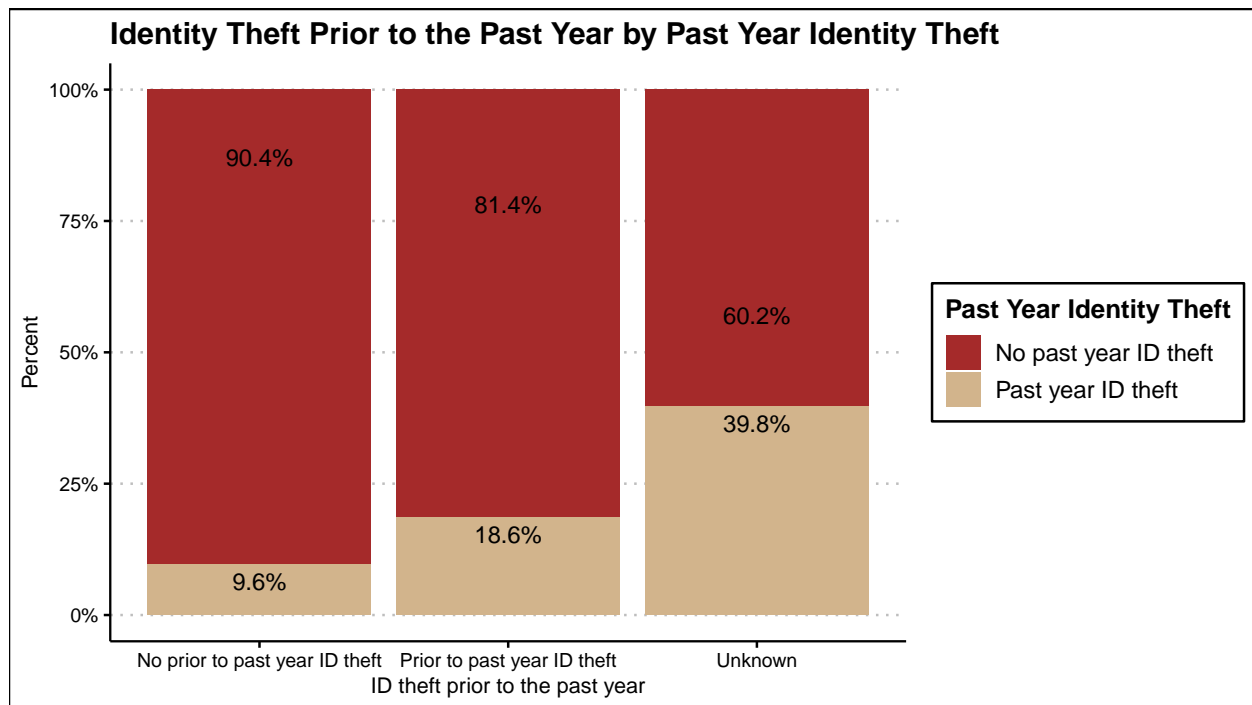
##	Count	Percentage
## Total	96130	100
## No	11066	12
## Unknown	368	0
## Yes	84696	88



### Identity theft prior to the past year

Thirteen percent of the sample experienced identity theft (misuse of an existing account, misuse of personal information to create new account or misuse of personal information for other fraudulent purposes) prior to 12 months prior to their ITS interview. The majority of the sample did not experience it. About 40% of respondents who did not know if they were a victim of identity theft prior to the past year experienced identity theft in the past 12 months. This is compared to 10% of those who had no identity theft prior to the past year and 19% of those who had identity theft prior to the past year.

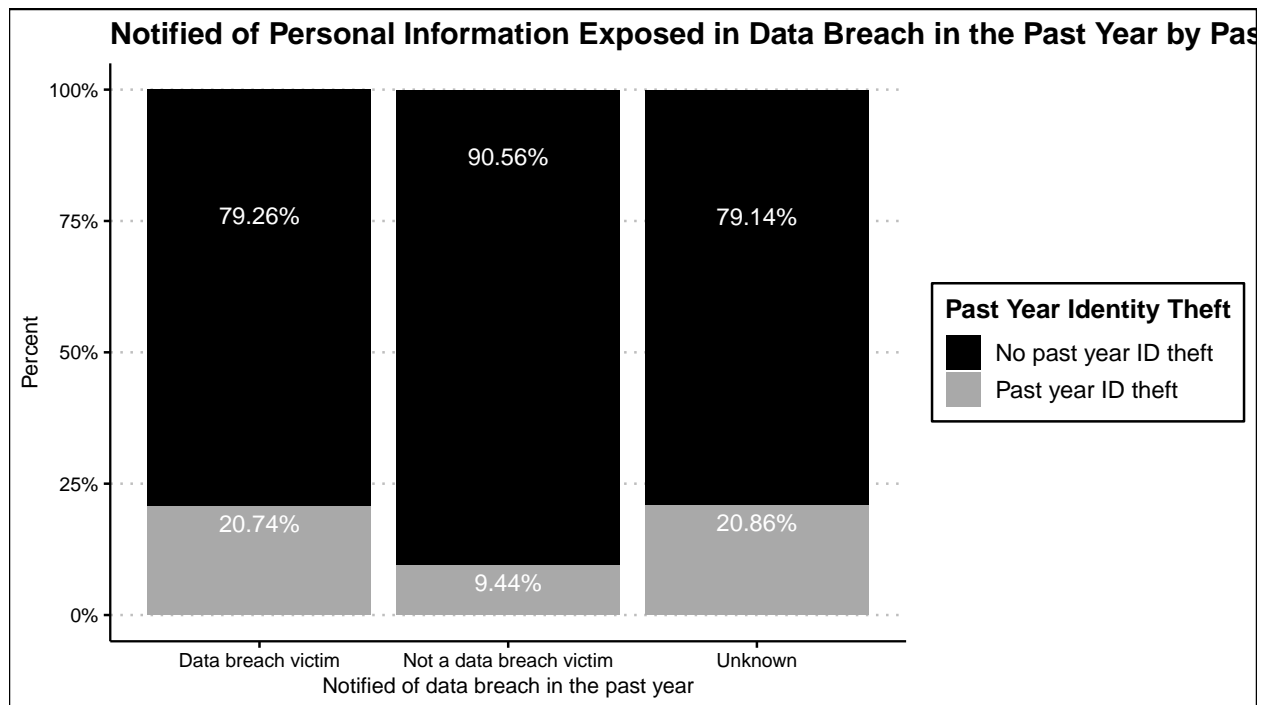
##	Count	Percentage
## Total	96130	100
## No prior to past year ID theft	83692	87
## Prior to past year ID theft	12267	13
## Unknown	171	0



## Notified of exposure due to data breach

Twelve percent of the sample reported that they were notified that their personal information was exposed during a data breach. The majority of the sample (88%) reported that they were not notified that their personal information was exposed during a data breach. Of those who were notified that their information was exposed during a data breach, one in five (21%) reported being victims of identity theft in the past year. This is compared to 9% of those who were not notified that their personal information was exposed in a data breach being victims of past year identity theft.

##	Count	Percentage
## Total	96130	100
## Data breach victim	11037	11
## Not a data breach victim	84652	88
## Unknown	441	0



## Data analysis

### More data wrangling

Make copies of each variable used in analysis. Unknown level on each individual variable was changed to NA. Individual variables were combined into a single dataset and deleted individual variables. The individual variables were combined into a dataset (its1). The number of complete (cases with no NA values on any variable) and incomplete cases (cases with at least one variable with a value of NA) was summed from the dataset. There were 614 cases with a value of NA on at least one variable with the remaining cases being complete cases

```
## 'data.frame': 96130 obs. of 8 variables:
## $ idtheft : Factor w/ 2 levels "No past year ID theft ",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ incomer : Factor w/ 2 levels "Less than $75,000",...: 2 2 2 2 1 1 1 1 1 1 ...
## $ ager : Factor w/ 6 levels "Age 16 to 17",...: 4 5 2 6 5 3 3 5 5 6 ...
## $ ethnicr : Factor w/ 2 levels "NonHispanic White",...: 1 2 2 1 1 1 1 1 1 1 ...
## $ prevent_total : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
## $ OUTSIDE_PAST_YEARR: Factor w/ 2 levels "No prior to past year ID theft",...: 1 1 2 1 1 2 1 1 1 1 .
## $ notify_breachr : Factor w/ 2 levels "Data breach victim",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ sexr : Factor w/ 2 levels "Male","Female": 2 1 1 1 1 2 1 1 2 2 ...

##          Number Percent
## Total          96130     100
## Cases with NAs         614        1
## Cases without NAs    95516     99
```

## Chi-Square Analysis

Multiple chi-square analyses were run on the dataset with only completed cases which meant that 614 cases were dropped from data analysis, leaving 95,516 cases to analyze. The chi-square analyses show that between past year identity theft was dependent on most of the predictors ( $p < 0.05$ ) with the exception of sex ( $p > 0.05$ ).

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: its1$idtheft and its1$incomer
## X-squared = 962.27, df = 1, p-value < 2.2e-16

##
## Pearson's Chi-squared test
##
## data: its1$idtheft and its1$ager
## X-squared = 544.24, df = 5, p-value < 2.2e-16

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: its1$idtheft and its1$sexr
## X-squared = 0.39041, df = 1, p-value = 0.5321

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: its1$idtheft and its1$ethnricr
## X-squared = 359.61, df = 1, p-value < 2.2e-16

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: its1$idtheft and its1$prevent_total
## X-squared = 1117.8, df = 1, p-value < 2.2e-16

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: its1$idtheft and its1$OUTSIDE_PAST_YEAR
## X-squared = 899.98, df = 1, p-value < 2.2e-16

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: its1$idtheft and its1$notify_breachr
## X-squared = 1291.3, df = 1, p-value < 2.2e-16
```