

# **Data Explorer 2.0: Analyse and Reuse data in your repository**

**Dataverse Workshop, April 17, 2023**

# Borealis (<https://borealisdata.ca>)

Canadian Dataverse  
Repository, provided in  
partnership with Canadian  
academic libraries.

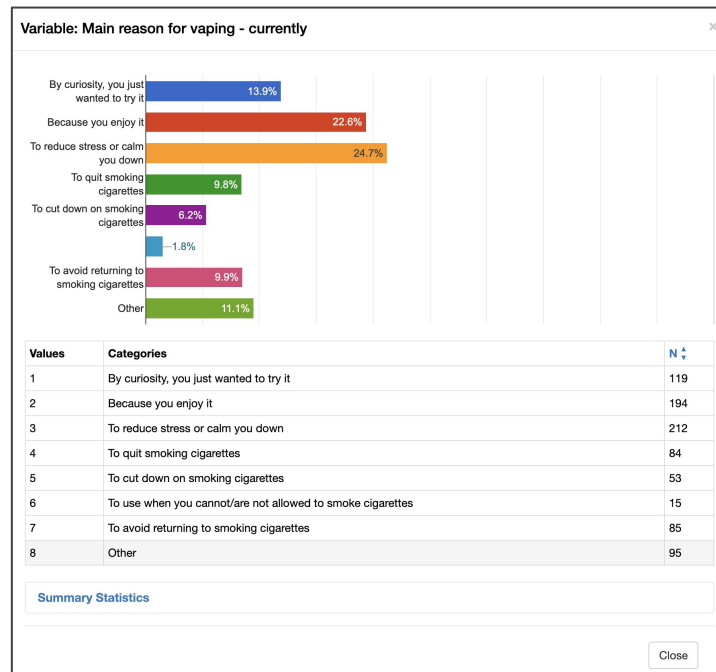
- Bilingual platform
- Institutional Collections
- Open repository development
- Shared standards

The screenshot displays the Borealis website interface. At the top, the Borealis logo is on the left, and navigation links for Search, User Guide, Support, English, and Log In are on the right. Below the header, a message states: "Borealis For testing Borealis, go to the demo site. // Pour tester Borealis, allez sur notre site de démonstration." A metrics bar shows "13,997,262 Downloads" and links for Contact and Share. A carousel of partner institutions is visible, including Algoma University, Brandon University, Brock University, and Cape Breton University. A search bar contains the text "Search this dataverse..." and "Advanced Search". On the left sidebar, filters are shown for "Dataverses (1,103)", "Datasets (13,669)", and "Files (235,901)". Under "Dataverse Category", there are links for Researcher (205), Organization or Institution (175), Research Group (123), and Laboratory (54). Under "Publication Year", there are links for 2023 (1,158), 2022 (5,201), 2021 (4,213), 2020 (891), and 2019 (1,195). The main content area shows "1 to 10 of 14,772 Results" with a sort dropdown. Three search results are displayed, each for a document by Dave Leip, dated Apr 12, 2023. The results are: "Governor Election Data", "US House Elections", and "Senate Election Data". Each result includes a DOI link and a brief description of the data.

# Data exploration

— — —

- Helps users understand the data
  - Preview data, documentation, codebooks, readme
  - Get started with data; digging, filtering, previewing
- Identify and find data
  - Variables and questions
  - Frequencies, statistics, cross-tabulation
  - Chart visualization
  - Subsetting & downloading
- Analyse and reuse training
  - Transparency and reproducibility
  - Teaching data concepts
  - Open access and sharing



Source: VAP\_35R, Canadian Tobacco and Nicotine Survey, 2021.  
Odesi <https://odesi.ca>

# Dataverse Repository

- Social science datasets are supported in Dataverse (many features)
  - Descriptive metadata
  - Variable metadata
  - Tabular data ingest (TAB file)
  - File level APIs
- Metadata Standards
  - Data Documentation Initiative (DDI)
  - Dataverse JSON
  - Other disciplinary connections
- Dataverse APIs
- R Tools
- Integrates with Data Explorer (1.0, 2.0)



Geospatial Metadata ▾	
Social Science and Humanities Metadata ▲	
<b>Unit of Analysis</b> ⓘ	Individuals
<b>Universe</b> ⓘ	The target population for the survey is non-institutionalized persons aged 15 years or older living in Canada's ten provinces, who are not members of collectives or living on reserves.
<b>Frequency</b> ⓘ	Occasional
<b>Sampling Procedure</b> ⓘ	Sampling unit: Individuals aged 15 to 24. The CTNS sample has a one-stage design, and the individual is the sampling unit. Individuals aged 25 and older. The CTNS sample has a two-stage design: the sampling unit for the first stage is the dwelling, and the sampling unit for the second stage is the individual. Stratification method: Individuals aged 15 to 24. The frame was stratified by age group (15 to 19, 20 to 24) and province, and a systematic sample was selected independently within each age group and province. Individuals aged 25 and older. The frame was stratified by province and a simple random sample of dwellings was selected independently within each province. Sampling and subsampling: Individuals aged 15 to 24. The sample was allocated to the provinces so that the survey could produce province-level estimates for this age group. The sample was also allocated so that the survey could produce region-level estimates for age groups 15 to 19 and 20 to 24. An initial sample of 12,000 individuals aged 15 to 24 was selected. Individuals aged 25 and older. The sample was allocated to the provinces so that the survey could produce province-level estimates for this age group. An initial sample of 12,000 dwellings was selected to cover individuals aged 25 and older. An initial sample of 24,000 dwellings or individuals was sent to collection.
<b>Collection Mode</b> ⓘ	Data are collected directly from survey respondents either through an electronic questionnaire (EQ) or through CATI (computer-assisted telephone interviewing).

# Data Explorer

---

- Dataverse Tabular file (.tab) & Data Explorer
- Preview tabular data
  - Variables & frequencies
- Summary statistics
  - (Averages, category frequencies, tabulations)
- Integrated external application since Dataverse version 4.6
- Data Explorer 1.0–2.0 in [Github](#)
  - Angular web framework
  - New in 2.0:
    - Variable Groups
    - View Question Text, Universes
    - Frequencies now include weights
    - Updated cross-tabulation feature
      - Multivariate analysis

# Canadian Tobacco and Nicotine Survey, 2021

English

CTNS-5305-E-2021\_F1.tab

Statistics Canada, 2023, "Canadian Tobacco and Nicotine Survey, 2021", <https://doi.org/10.5072/FK2/ZHT5KX>, Borealis, V1, UNF:6:I5CjFBAeehXclm89HK6dpA== [fileUNF]

< Hide Groups

 Cross Tabulation

 Summary Statistics

Download

All Variables








































Search



Items per page 25

1 - 25 of 104



	<input type="checkbox"/>	ID	Name	Label	Weight	View Summary Statistics	View Categories	View Questions
CAN: Cannabis	<input type="checkbox"/>							
DV: Derived Variables	<input type="checkbox"/>	v6423547	AGEGROUP	Age group of person				
ALC: Alcohol	<input type="checkbox"/>	v6423552	ALC_05	Frequency drank alcoholic beverage - past 30 days				
IU: Initial use	<input type="checkbox"/>	v6423564	ALC_10	Freq drank 4 drinks or more on one occasion - past 12 months				
VAP: Vaping	<input type="checkbox"/>	v6423571	CAN_05AR	Smoked cannabis - lifetime				
OTP: Other tobacco product status	<input type="checkbox"/>	v6423570	CAN_05BR	Age first time smoked cannabis				
RD: Reference date	<input type="checkbox"/>	v6423525	CAN_10AR	Frequency smoked cannabis - past 30 days				
DEM2: Demographics 2	<input type="checkbox"/>	v6423539	CAN_10BR	Nb of days smoked cannabis at least once past month - grouped				
SRV: Survey Related Variables	<input type="checkbox"/>	v6423524	CAN_15AR	Frequency smoked cannabis/tobacco mix - past 30 days				
GDR: Gender	<input type="checkbox"/>	v6423569	CAN_15BR	Nb of days smoked can./toba. mix at least once a week - past 30 days				
TBC: Tobacco	<input type="checkbox"/>	v6423529	CAN_17R	Frequency consumed edibles - past 30 days				
	<input type="checkbox"/>	v6423486	CAN_20AR	Vaped cannabis - lifetime				
	<input type="checkbox"/>	v6423483	CAN_20BR	Age first time vaped cannabis - grouped				
	<input type="checkbox"/>	v6423541	CAN_25AR	Frequency vaped cannabis - past 30 days				

# Data Explorer tools

---



# Depositing and submitting data

- Upload tabular files (SPSS, STATA, SAS, R, CSV, Excel)
- Data ingest tips:
  - Responses/ observations as rows
  - Variables/concepts as columns
  - No notes / improper formatting
  - Does not support multiple spreadsheets in one file
  - Excel (XLST) is supported
  - CSV requires documentation
- [Guides](#) and other resources

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
1	1.040e+14	4	26	104012	99	1	1973	2	1642	1040	1040	1	6	1	5	5
2	1.040e+14	1	19	104008	99	2	1943	1	99	99	99	1	1	6	4	5
3	1.040e+14	4	98	104003	99	1	1990	1	99	99	99	1	6	3	99	11
4	1.040e+14	2	98	104010	99	1	1983	2	1756	1040	1040	1	6	3	99	11
5	1.040e+14	1	18	104007	99	2	1927	1	99	99	99	1	1	6	4	1
6	1.040e+14	2	19	104007	99	1	1983	1	99	99	99	1	6	2	5	3
7	1.040e+14	4	15	104005	99	2	1970	1	99	99	99	1	1	2	4	6
8	1.040e+14	4	19	104006	99	1	1942	1	99	99	99	2	1	6	5	88
9	1.040e+14	4	15	104005	99	1	1965	2	1040	1276	1040	7	4	1	5	5
10	1.040e+14	7	15	104004	99	2	1955	1	99	99	99	1	1	2	5	5
11	1.040e+14	7	15	104003	99	2	7777	1	99	99	99	1	1	6	77	77
12	1.040e+14	4	18	104005	99	2	1938	1	99	99	99	1	3	6	5	5
13	1.040e+14	7	17	104005	99	1	1945	1	99	99	99	1	1	6	5	4
14	1.040e+14	4	18	104005	99	2	1949	2	1040	1300	1040	1	4	6	5	5
15	1.040e+14	2	15	104003	99	2	7777	1	99	99	99	1	1	2	4	5
16	1.040e+14	4	32	104007	99	1	1974	1	99	99	99	1	6	2	5	5
17	1.040e+14	4	98	104006	99	1	7777	1	99	99	99	1	6	3	99	11
18	1.040e+14	4	25	104010	99	2	1968	1	99	99	99	1	1	2	5	1
19	1.040e+14	4	40	104007	99	1	1967	1	99	99	99	1	1	2	6	2
20	1.040e+14	1	23	104004	99	1	1932	1	99	99	99	1	1	6	4	3
21	1.040e+14	2	18	104003	99	2	1965	1	99	99	99	1	6	2	5	5
22	1.040e+14	1	27	104011	99	1	1956	1	99	99	99	2	1	1	5	2
23	1.040e+14	4	18	104004	99	2	1923	1	99	99	99	1	3	6	3	3
24	1.040e+14	1	19	104006	99	2	1952	1	99	99	99	1	4	2	5	3
25	1.040e+14	3	16	104004	99	2	1947	2	1276	1040	1040	1	1	6	3	3



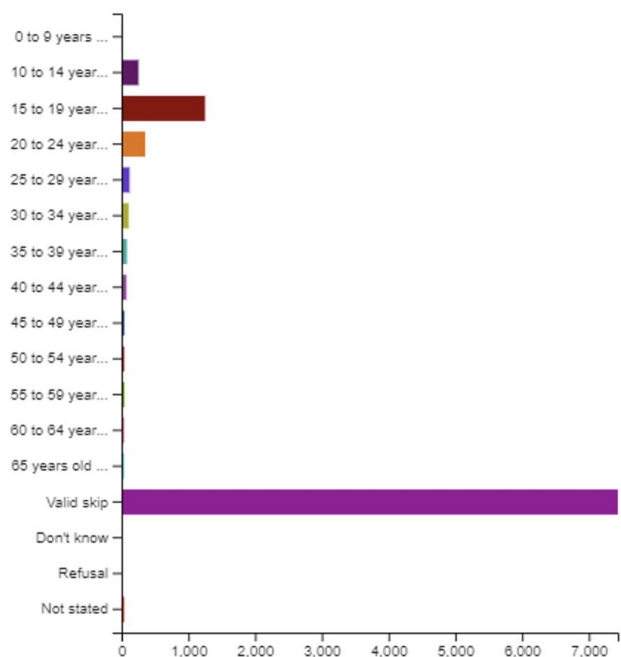
# Variable metadata

- Assigns Unique identifiers
- Names, labels, codes, values
- Description of content
  - Interviews, question text, universes, recodes and notes
- Weighting
- Summary statistics
- Export HTML and PDF Codebooks
- Export XML metadata for integrations

Variable Description
<b>Variable Groups</b>
<ul style="list-style-type: none"><li>• <a href="#">CAN: Cannabis</a></li><li>• <a href="#">DV: Derived Variables</a></li><li>• <a href="#">ALC: Alcohol</a></li><li>• <a href="#">IU: Initial use</a></li><li>• <a href="#">VAP: Vaping</a></li><li>• <a href="#">OTP: Other tobacco product status</a></li><li>• <a href="#">RD: Reference date</a></li><li>• <a href="#">DEM2: Demographics 2</a></li><li>• <a href="#">SRV: Survey Related Variables</a></li><li>• <a href="#">GDR: Gender</a></li><li>• <a href="#">TBC: Tobacco</a></li></ul>
<b>CAN: Cannabis</b>
Variables within CAN: Cannabis
<ul style="list-style-type: none"><li>• <a href="#">Smoked cannabis - lifetime</a></li><li>• <a href="#">Age first time smoked cannabis</a></li><li>• <a href="#">Frequency smoked cannabis - past 30 days</a></li><li>• <a href="#">Nb of days smoked cannabis at least once past month - grouped</a></li><li>• <a href="#">Frequency smoked cannabis/tobacco mix - past 30 days</a></li><li>• <a href="#">Nb of days smoked can /toba mix at least once a week - past 30 days</a></li><li>• <a href="#">Frequency consumed edibles - past 30 days</a></li><li>• <a href="#">Vaped cannabis - lifetime</a></li><li>• <a href="#">Age first time vaped cannabis - grouped</a></li><li>• <a href="#">Frequency vaped cannabis - past 30 days</a></li><li>• <a href="#">Nb of days vaped cannabis at least once past month - 30 days - grouped</a></li><li>• <a href="#">Where get devices/liquids - Make your own vaping liquid</a></li><li>• <a href="#">Where get devices/liquids - Compassion club/dispensary/storefront</a></li><li>• <a href="#">Where get devices/liquids - Online</a></li><li>• <a href="#">Where get devices/liquids - Shared around a group of friends</a></li><li>• <a href="#">Where get devices/liquids - An acquaintance</a></li><li>• <a href="#">Where get devices/liquids - A family member</a></li><li>• <a href="#">Where get devices/liquids - A friend</a></li><li>• <a href="#">Where get devices/liquids - A dealer</a></li><li>• <a href="#">Where get devices/liquids - Other</a></li></ul>
<b>DV: Derived Variables</b>
Variables within DV: Derived Variables
<ul style="list-style-type: none"><li>• <a href="#">Age group of person</a></li><li>• <a href="#">Province of residence (collection)</a></li><li>• <a href="#">Smoking Status</a></li><li>• <a href="#">Cannabis smoked in the last 30 days</a></li><li>• <a href="#">Vaped in the last 30 days</a></li><li>• <a href="#">Vaped cannabis in the last 30 days</a></li><li>• <a href="#">Consumed alcohol in the last 30 days</a></li><li>• <a href="#">First product tried</a></li></ul>

Source: Canadian Tobacco and Nicotine Survey, 2021.  
DDI Codebook (Odesi)

## VAP\_05BR: Age first time tried vaping ×



Values	Categories	Count	Count Percentage(%)	Weighted Count	Weighted Count Percentage(%)
1	0 to 9 years old	5	0.05		
2	10 to 14 years old	253	2.553		

## Variable Information ×

**ID:** v6423478

**Name:** VAP\_05BR

**Label:** Age first time tried vaping

**Literal Question:** How old were you when you first tried vaping?

**Universe:** VAP\_05AR = 1

**Weight Variable:** Unweighted

**Is Weight:** No

# Editing variable metadata

---

- Variable metadata is ingested and stored in Dataverse database
- Related [Data Curation Tool](#) supports editing variable metadata
  - Add variable groups, question/concept description, universe notes, apply weights
- Edits are saved as new version in DV
- Updates are displayed in Data Explorer

The screenshot displays the 'Variable Information' form in the Data Curation Tool. The form is overlaid on a background showing a list of variables. The 'Variable Information' form includes the following fields:

- ID: v29744
- Name: AGEGRP
- Label: Groupe d'âge du répondant
- Literal Question: (empty)
- Interviewer Instructions: (empty)
- Post Question: (empty)
- Universe: (empty)
- Notes: (empty)
- Group: (empty)

The background shows a table of variables with the following entries:

Name
PUMFID
VERDATE
AGEGRP
BH_05
BH_10
BH_25
BH_30
BH_45
PBH_50

# Connecting to Data Explorer

odesi  Français

Find data:

Enter search term

Anywhere

+ Add row

Collection:

All

Sort:

Relevance

Date (YYYY):

From

To

Include:  Research datasets

SEARCH

Odesi  
for the  
than 7  
descri

## Canadian Tobacco and Nicotine Survey, 2020

EXPLORE VARIABLES

ASSOCIATED FILES

URI [odesi.ca/uri/odesi/doi\\_10-5072\\_FK2\\_1URBO8.xml](https://odesi.ca/uri/odesi/doi_10-5072_FK2_1URBO8.xml)

The main objective of the Canadian Tobacco and Nicotine Survey is to gather information about the prevalence of cigarette smoking, vaping, and cannabis use. Until 2017, Statistics Canada conducted the Canadian Tobacco, Alcohol and Drug Survey (CTADS), which collected data on tobacco as well as alcohol and drug use in Canada. In 2020 the Canadian Alcohol and Drug Survey (CADS) was conducted to collect data on alcohol and drug use independently from the Canadian Tobacco and Nicotine Survey (CTNS) which was conducted to primarily collect data on tobacco and nicotine.

**Citation** Statistics Canada, 2022, "Canadian Tobacco and Nicotine Survey, 2020", <https://doi.org/10.5072/FK2/1URBO8>, Borealis, V1, UNF:6:m2tmN0jECi8+h0latlN9g== [fileUNF]

**Authoring Entity** Statistics Canada

**Geographic Coverage** Province

**Geographic Unit** Canada



The screenshot shows the Odesi Data Explorer interface. On the left, there is a search bar and filters. The main area displays a list of files:

- CTNS-5305-E-2021\_F1.tab**: Tabular Data - 2.4 MB, Published Apr 15, 2023, 0 Downloads, 104 Variables, 9908 Observations UNF:6:i5Cj...dpA==, Microdata in tab separated value format.
- CTNS2021\_AbridgedUserGuide.pdf**: Survey documentation/ Adobe PDF - 929.6 KB, Published Apr 15, 2023, 0 Downloads, MDS: 717...891, Data Documentation.
- CTNS2021\_PUMF\_Cdbk.pdf**: Survey documentation/ Adobe PDF - 1.4 MB, Published Apr 15, 2023, 0 Downloads, MDS: fa9...6f4, Data Documentation.
- CTNS2021\_Questionnaire\_E.pdf**: Survey documentation/ Adobe PDF - 741.6 KB, Published Apr 15, 2023, 0 Downloads.

A dropdown menu is open on the right, showing options: File Access, Public, Download Options, SPSS Binary (Original File Format), Tab-Delimited, RData, Download Metadata, Variable Metadata, Data File Citation, Explore Options, View Data, and **Data Explorer** (highlighted with a red box).

# Preview data

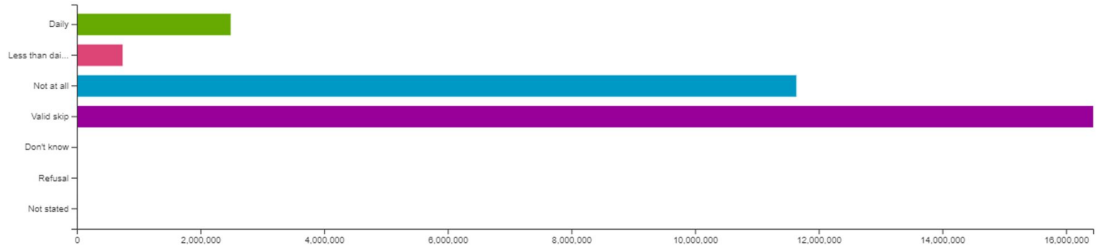
smoking policies on smoking behaviour, the Canadian Tobacco Use Monitoring

EXPLORE VARIABLES 

MATCHING VARIABLES 96

ID	Name	Label
1	v2609851	Any in hhd smoke cigarettes
2	v2609786	# of people that smoke inside the home
3	v2609819	# of cigs. smoked inside home
4	v2609923	Is smoking cigs. allowed in your home
5	v2609850	Is smoking cigs. restricted in your home
6	v2609818	Smoking restricted-in certain rooms only
7	v2609814	Smoking restricted-presence of children
8	v2609808	Smoking restricted-windows open
9	v2609789	Smoking restricted-other
10	v2609815	Smoke cigs. daily/occasional/not at all
11	v2609877	Past 30 days, smoked any cigarettes
12	v2609853	Past 30 days, smoke every day
13	v2609874	Smoked at least 100 cigarettes in life
14	v2609898	Smoked a whole cigarette
15	v2609845	Smoked cigarettes daily
16	v2609839	Age smoked first cigarette

Variable: Frequency smoked cigarettes - past 30 days



Values	Categories	Count	Count Percentage(%)	Weight Count
1	Daily	608	7.495	2487936.268
2	Less than daily, but at least once in the past month	184	2.268	740806.37
4	Not at all	2852	35.158	1.1637711838E7
6	Valid skip	4464	55.030	1.6439278195E7
7	Don't know	0	0.000	0.0
8	Refusal	0	0.000	0.0
9	Not stated	4	0.049	890.84

# Analyse data

## Select Variables For Cross Tabulation

- VAP\_15AR - Nb of days vaped e-liquid with nicotine - pas
- VAP\_15BR - Nb of days vaped e-liquid without nicotine -
- VAP\_15CR - Nb of days vaped e-liquid didn't know what it
- VAP\_20R - Nb of times pick up vaping device - Day respo
- VAP\_21R - Nb of puffs taken each time vaping device is
- VAP\_30R - Flavour respondent vape most often
- VAP\_40AR - Where get vaping devices - Vape shop
- VAP\_40BR - Where get vaping devices - Convenience store/
- VAP\_40CR - Where get vaping devices - Supermarket/grocer
- VAP\_40DR - Where get vaping devices - Online



Rows  
 VAP\_35R - Main reason for vaping - currently



Columns  
 AGEGROUP - Age group of person



## Cross Tabulation Table

AGEGROUP - Age group of person	15 to 19 years old	20 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 years old and older	Valid skip	Don't know	Refusal	Not stated	Total
<b>VAP_35R - Main reason for vaping - currently</b>												
By curiosity, you just wanted to try it	59 0.6%	45 0.45%	7 0.07%	3 0.03%	3 0.03%	1 0.01%	1 0.01%	0 0%	0 0%	0 0%	0 0%	119 1.2%
Because you enjoy it	72 0.73%	102 1.03%	9 0.09%	7 0.07%	2 0.02%	1 0.01%	1 0.01%	0 0%	0 0%	0 0%	0 0%	194 1.96%
To reduce stress or calm you down	87 0.88%	105 1.06%	7 0.07%	8 0.08%	2 0.02%	0 0%	3 0.03%	0 0%	0 0%	0 0%	0 0%	212 2.14%
To quit smoking cigarettes	6 0.06%	37 0.37%	10 0.1%	11 0.11%	12 0.12%	2 0.02%	6 0.06%	0 0%	0 0%	0 0%	0 0%	84 0.85%
To cut down on smoking cigarettes	1 0.01%	28 0.28%	4 0.04%	8 0.08%	3 0.03%	5 0.05%	4 0.04%	0 0%	0 0%	0 0%	0 0%	53 0.53%
To use when you cannot/are not allowed to smoke cigarettes	2 0.02%	8 0.08%	0 0%	2 0.02%	2 0.02%	1 0.01%	0 0%	0 0%	0 0%	0 0%	0 0%	15 0.15%
To avoid returning to smoking cigarettes	2 0.02%	41 0.41%	15 0.15%	8 0.08%	12 0.12%	3 0.03%	4 0.04%	0 0%	0 0%	0 0%	0 0%	85 0.86%
Other	31 0.31%	49 0.49%	3 0.03%	6 0.06%	2 0.02%	2 0.02%	2 0.02%	0 0%	0 0%	0 0%	0 0%	95 0.96%
	1775	1087	641	612	841	1100	1764	0	0	0	0	6070

# Download and subset

Canadian Tobacco and Nicotine Survey, 2021 English ▾

CTNS-5305-E-2021\_F1.tab

Statistics Canada, 2023, "Canadian Tobacco and Nicotine Survey, 2021", <https://doi.org/10.5072/FK2/ZHT5KX>, Borealis, V1, UNF:6:15CjFBAeehXclm89HK6dpA== [fileUNF]

< Hide Groups Cross Tabulation Summary

Download

Download subset

Download original file

Download tab-delimited file

Download RData format file

Download variable metadata

All Variables  Items per page 25 ▾

	ID	Name	Label	Weight	
CAN: Cannabis					
DV: Derived Variables					
ALC: Alcohol					
IU: Initial use					
VAP: Vaping					
OTP: Other tobacco product status					
RD: Reference date					
DEM2: Demographics 2					
SRV: Survey Related Variables					
GDR: Gender					
TBC: Tobacco					
<input type="checkbox"/>	v6423547	AGEGROUP	Age group of person		
<input checked="" type="checkbox"/>	v6423552	ALC_05	Frequency drank alcoholic beverage - past 30 days		
<input checked="" type="checkbox"/>	v6423564	ALC_10	Freq drank 4 drinks or more on one occasion - past 12 months		
<input checked="" type="checkbox"/>	v6423571	CAN_05AR	Smoked cannabis - lifetime		
<input checked="" type="checkbox"/>	v6423570	CAN_05BR	Age first time smoked cannabis		
<input type="checkbox"/>	v6423525	CAN_10AR	Frequency smoked cannabis - past 30 days		
<input type="checkbox"/>	v6423539	CAN_10BR	Nb of days smoked cannabis at least once past month - grouped		
<input type="checkbox"/>	v6423524	CAN_15AR	Frequency smoked cannabis/tobacco mix - past 30 days		
<input type="checkbox"/>	v6423569	CAN_15BR	Nb of days smoked can./toba. mix at least once a week - past 30 days		
<input type="checkbox"/>	v6423529	CAN_17R	Frequency consumed edibles - past 30 days		

# Examples & demonstration

— — —

Go to: `Odesi.ca`

[“Vaping”](#)

Canadian Tobacco and Nicotine Survey, 2021

[“Justin Trudeau”](#)

Canadian Election Study, 2021



# Conclusion

---

- Depositing and sharing tabular data in Dataverse can support data exploration
- Using Dataverse APIs you can build data exploration tools
- Data Explorer is an open, ready to use tool
- Data Explorer can be built into integrated data search sites, such as Odesi.ca

For more information about Data Explorer reach out to us at [borealisdata.ca](http://borealisdata.ca) or contact me at [amber.leahey@utoronto.ca](mailto:amber.leahey@utoronto.ca)