Statement of Intent

I am choosing Government of Canada's open data resources for data collection of Tax-Free Savings Accounts for last 4 years to perform Analysis for my Capstone Project and my career aspiration is to become a Data Analyst with any of the Federal Government Departments.

My motivation for pursuing research related to data analysis stems from learning about the challenges of software development applications. Although I found the actual development work to be educational, what really sparked my interest in research was experiencing first-hand the challenges that software engineers face daily in researching, evaluating, analyzing, and implementing new IM/IT technologies, trends, and practices.

https://search.open.canada.ca/en/od/?od-search-portal=Open%20Data&search_text=tax+free#

Data Analysis Proposal

1) Project Title:

The Tax-Free Savings Account (TFSA) statistics tables present data based on all the individual records and summary returns sent by financial institutions (issuers) to the CRA for the 2017 tax year and processed up to January 21, 2019. Only the most current information is considered valid for data purposes.

The data in the tables are taken from the tax year 2014 to 2017 which is typically published two years after the tax year ends.

2) Rationale and objectives of the study:

To ensure the protection of taxpayer information, data have been suppressed where warranted. As well, counts are rounded to the nearest multiple of 10.

Income data were taken from income tax returns and related schedules filed by individuals for the 2014 to 2017 tax year.

This project is based on Descriptive Analysis; hence statistics provide absolute numbers. However, they do not explain the rationale or reasoning behind those numbers. In order to that higher access to data must be needed. Before applying descriptive statistics, it's important to think about which one is best suited for your research question and what you want to show. For example, a percentage is a good way to show the age distribution of respondents in various provinces. Typically, descriptive statistics is the first level of analysis. It helps researchers summarize the data and find patterns.

Research Objectives:

The purpose of this project is to find absolute numbers for the following questions:

- Which Province has maximum increase of TFSA account holders in last 4 years.

- Which Province has maximum increase of TFSA account holders in last 4 years with respect to age and income class.
- Which Province has maximum and minimum TFSA Fair Market Value, Contributions and Withdrawals in last 4 years.
- Which Province has maximum and minimum TFSA Fair Market Value, Contributions and Withdrawals with respect to age and income class in last 4 years.

3. Proposed data analysis and software requirements:

The data analysis will be carried out using Power BI visualization package and its integrated components.

The data analysis will involve two steps. The first is descriptive statistics represented by bar charts and contingency tables to measure the relationship between different independent variables. A few commonly used descriptive statistics are:

- Mean: numerical average of a set of values.
- Median: midpoint of a set of numerical values.
- Mode: most common value among a set of values.
- Percentage: used to express how a value or group of respondents within the data relates to a larger group of respondents.
- Frequency: the number of times a value is found.
- Range: the highest and lowest value in a set of values.

Descriptive statistics are most helpful when the research is limited to the sample and does not need to be generalized to a larger population. For example, if you are comparing the percentage of Fair Market Value in two different Provinces, then descriptive statistics is enough.

Since descriptive analysis is mostly used for analyzing single variable, it is often called univariate analysis.

The second step in the data analysis is regression modeling. Regression analysis is primarily used for two conceptually distinct purposes. First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Second, in some situation's regression analysis can be used to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset. To use regressions for prediction or to infer causal relationships, respectively, a researcher must carefully justify why existing relationships have predictive power for a new context or why a relationship between two variables has a causal interpretation. The latter is especially important when researchers hope to estimate causal relationships using observational data

4) Data Requirements:

At this point of research, the data is available on Open Government of Canada Portal and no access is required for any of the confidential Master Data Files from Canada Revenue Agency, however access can be requested in the higher stages of analysis. There are three major classification variables described in TFSA tables.

Major Classification Variables:

The following variables are used in one or more of the tables in this publication:

- Age
- province or territory of residence
- total income

Age classification:

The TFSA holder's age is determined using the reported year of birth on page 1 of the T1 General Income Tax and Benefit Return. Individuals with no reported date of birth are included in the total.

Provincial or territorial classification:

Province or territory of residence - used in tables 1B and 3B - refers to the province or territory in which the tax filer resided on December 31, 2014, as indicated in the T1 General Income Tax and Benefit Return. If the tax filer province or territory of residence is missing or unclear, they are classified as "Other".

Total income classification:

Total income class – used in tables 1C and 3C – refers to the 'total income assessed' reported on Line 150 of the T1 General Income Tax and Benefit Return in the 2014 tax year.

Description of TFSA tables:

Each table contains the number of TFSAs, TFSA holders or the respective dollar amounts. In some cases, the total of the figures in the table may not match the total shown due either to rounding or to editing for confidentiality purposes.

Tables 1, 1A, 1B, 3, 3A and 3B:

Tables 1A and 3A present information by age group, ranging from under-20 to 75-and-over. The grand total includes tax filers whose age is not stated.

Tables 1B and 3B present information according to the province or territory of residence listed on the tax filer's income tax return.

Tables 1C and 3C:

Tables 1C and 3C present 21 income groups based on total income assessed, ranging from loss and nil to $250,000 and over. The ranges include an "N/A" group that represents a segment of Canadians for which CRA has TFSA data but no income tax data. It should be noted that Canadians are not required to file an income tax return to open or to use a TFSA.

Certain types of income are not included in total income assessed because they are non-taxable, so true economic income may be understated. An overstatement may be caused by other types of income that are grossed-up (such as eligible dividends grossed-up to 138%) or gross income. For a description of the income components

Table 2:

This table presents information on the number of TFSAs per tax filer.