Analysis of non-profit-organization in Canada before 2021

Order for use with the data

- Data preparation dataset - data_analysis_categorized_technical_report.ipynb or .py
- Data preparation only - data_analysis_categorized_...-data_preparation_only.py
- Data Analysis - Final_Reslt/Portion_Tehcnial_Report_Final.ipynb or .py
- Data Analysis Testing only - Final_Reslt/Portion_Tehcnial_Report_Final_Select.ipynb or .py
- Report and Slide - Document-Report/

Process of analysis

- Import CSV file into the one big dataset.
- Filtered some columns and attributes from the dataset.
- Removed null values from the dataset.
- Divide into four different datasets based on the year
  - Contain three years worth of the data
  - Datasets splited into 2010-2012, 2013-2015, 2016-2018, 2019-2021
- Combined into two different datasets, training and testing set.
  - Training set, 2013-2018 (Combined 2013-2015, 2016-2018), 60-65%
  - Testing set, 2019-2021, 40-45%
  - Unused '2010-2012' dataset will be kepted for backup.
- Division into four (total of eight) different datasets from column called 'characteristics'.
  - Training set, four different datasets
  - Testing set, four different datasets
  - Remaining unused 'characteristics' will be dropped
- Division based on column called 'GEO' based on provinces. There will be thirteen 'GEO' data.
  - Training set, 4*13 different datasets.
    - Provinces, 13 different datasets.
  - Testing set, 4*13 different datasets
    - Provinces, 13 different datasets.
- There will be four (total of eight) different datasets by selected 'five' provinces and merged from previous four (or eight) datasets.
  - Training set, four different datasets with five provinces added
  - Testing set, four different datasets with five provinces added
  - Remaining unused 'GEO/provinces' will be dropped
- Five provinces added will be represented as binary (one-hand encoding) and characteristics values will be represented as numeric values.

Variable names involve during the analysis

- df - Whole dataset without any filtering or division
- df_sorted - Whole dataset with any filtering like removing non-important attributes.
- df_sorted_na - Whole dataset with removal of the null values inside the dataset.

Division of into new dataset based on Indicator

- df_AvgAnnHrsWrk - Average annual hours worked
- df_AvgAnnWages - Average annual wages and salaries
- df_AvgHrsWages - Average hourly wage
- df_AvgWeekHrsWrked - Average weekly hours worked
- df_Hrs_Wrked - Hours Worked
- df_NumOfJob - Number of jobs
- df_WagesAndSalaries - Wages and Salaries

Division of into new dataset based on the GEO/year

- df_AvgAnnHrsWrk_2010 - Average annual hours worked in 2010
- df_AvgAnnHrsWrk_2013 - Average annual hours worked in 2013
- df_AvgAnnHrsWrk_2016 - Average annual hours worked in 2016
- df_AvgAnnHrsWrk_2019 - Average annual hours worked in 2019

Then merge into

- training_df_AvgAnnHrsWrk - Average annual hours worked for training set (2013-2018)
- testing_df_AvgAnnHrsWrk - Average annual hours worked for testing set (2019-2021)

When splited by Characteristics type

- testing_df_AvgAnnHrsWrk_ByAge - Average annal hours worked by age group (Testing set)
- testing_df_AvgAnnHrsWrk_ByGender - Average annual hours worked by gender type (Testing set)
- testing_df_AvgAnnHrsWrk_ByEducation - Average annual hours worked by education level (Testing set)
- testing_df_AvgAnnHrsWrk_ByImmigrant - Average annual hours worked by immigrant group (Testing set)

When splitied by Provinces

- testing_df_AvgAnnHrsWork_ByAge_Provinces - Average annual hours worked by age group for all provinces (Testing set)
- testing_df_AvgAnnHrsWrk_ByAge_FiveProvinces - Average annual hours worked by age group for five provinces with analysis (Testing set)