Introduction

The United States and Canada share a rich cultural, historical, and geographical similarities that bind the two nations together. Both countries boast diverse populations, ethnicities and traditions, which contribute to their vibrant societies. Additionally, both countries have seen rising healthcare costs and increasing rates of chronic diseases, prompting discussions about healthcare and the need for improved public health initiatives. Given the similarities, we will examine whether the data on cause of deaths between the 2 countries would be in similar fashion. The goal of this project is to answer the above assumption along with analysis of common patterns and discrepancies. Python programming language will be utilized.

Data source

- The US: Publicly available files will be used. The first contains data on causes of death: "NCHS_-_Leading_Causes_of_Death__United_States.csv"; The second contains population data: "nst-est2018-01.xlsx" Both files have state-level information for multiple years.
- Canada: Publicly available data from Canadian government Canada.ca website: "13100394.csv"; (source: https://ouvert.canada.ca/data/dataset/99993095-becb-454b-9568-e36ae631824e)

Revisit US data

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.ticker import FuncFormatter
path = "/Users/qmacstore/Downloads/[BANA 680]/_Assignment/"
file = 'NCHS_-_Leading_Causes_of_Death__United_States.csv'
df = pd.read_csv(path+file)
# The likelihood in the US
df1 = df[(df['State']=='United States') & (df['Cause Name'] != 'All causes')]
df1a = df1.groupby(['Year'])['Deaths'].sum().reset_index()
df1b = df1a.T
df1b
```

Out[1]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Deaths	1905826	1902194	1899358	1918873	1912115	1864133	1889981	1854676	1846301	1872981	1838501	1852349	1869321	1876588

In [2]:
```
# What are the 4 leading causes of death for Americans?
df2 = df[(df['State']=='United States') & (df['Cause Name'] != 'All causes')]
df2b = df2.groupby(['Cause Name'])['Deaths'].sum()
sorted_grouped = df2b.sort_values(ascending=False)
top_4 = sorted_grouped.nlargest(4)
top_4
```

Out[2]:
```
Cause Name
Heart disease    11575183
Cancer           10244536
Stroke            2580140
CLRD              2434726
Name: Deaths, dtype: int64
```

In [3]:
```
# Are there year-by-year changes in the four leading causes of death nationwide?
us_deaths_causes_dfyr = df[(df['State'] == 'United States') & (df['Cause Name'] != 'All causes')]
rows = []
for yr in us_deaths_causes_dfyr['Year'].unique():
    state_cause_yrdf = us_deaths_causes_dfyr[us_deaths_causes_dfyr['Year'] == yr].groupby(['Cause Name'])['Deaths
    state_cause_yrdf.sort_values(ascending=False, inplace=True)
    cause_list = list(state_cause_yrdf.keys())[:4] + [None] * (4 - len(state_cause_yrdf))
    rows.append({'Year': yr, '1st cause': cause_list[0], '2nd cause': cause_list[1],
                 '3rd cause': cause_list[2], '4th cause': cause_list[3]})
df4 = pd.DataFrame(rows)
df4.sort_values(by='Year', ascending=True, inplace=True)
df4T = df4.T
df4T
```

Out[3]:

	1	0	2	7	3	6	4	5	8	9	10	11	12	14	13
Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
1st cause	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease	Heart disease

2nd cause	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer
3rd cause	Stroke	Stroke	Stroke	Stroke	Stroke	Stroke	Stroke	Stroke	Stroke	CLRD	CLRD	CLRD	CLRD	CLRD	CLRD
4th cause	CLRD	CLRD	CLRD	CLRD	CLRD	CLRD	CLRD	CLRD	CLRD	Stroke	Stroke	Stroke	Stroke	Stroke	Unintentional injuries

Canada

In [11]:
```
path1 = "/Users/qmacstore/Downloads/[BANA 680]/_FINAL/"
file1 = '13100394.csv'
df_ca = pd.read_csv(path1+file1)
df_ca.head(3)
```

```
<ipython-input-11-f9e2d84c4d57>:3: DtypeWarning: Columns (14) have mixed types. Specify dtype option on import or set low_memory=False.
  df_ca = pd.read_csv(path1+file1)
```

Out[11]:

	REF_DATE	GEO	DGUID	Age at time of death	Sex	Leading causes of death (ICD-10)	Characteristics	UOM	UOM_ID	SCALAR_FACTOR	SCALAR_ID	VECT
0	2000	Canada, place of residence	2016A000011124	Age at time of death, all ages	Both sexes	Total, all causes of death [A00-Y89]	Number of deaths	Number	223	units	0	v41618
1	2000	Canada, place of residence	2016A000011124	Age at time of death, all ages	Both sexes	Total, all causes of death [A00-Y89]	Percentage of deaths	Percentage	242	units	0	v41618
2	2000	Canada, place of residence	2016A000011124	Age at time of death, all ages	Both sexes	Total, all causes of death [A00-Y89]	Age-specific mortality rate per 100,000 popula...	Number	223	units	0	v41618

Exploratory data analysis (EDA)

In [10]:
```
df_ca.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 407334 entries, 0 to 407333
Data columns (total 18 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   REF_DATE                        407334 non-null  int64  
 1   GEO                             407334 non-null  object 
 2   DGUID                           407334 non-null  object 
 3   Age at time of death            407334 non-null  object 
 4   Sex                             407334 non-null  object 
 5   Leading causes of death (ICD-10)  407334 non-null  object 
 6   Characteristics                 407334 non-null  object 
 7   UOM                             407334 non-null  object 
 8   UOM_ID                          407334 non-null  int64  
 9   SCALAR_FACTOR                   407334 non-null  object 
 10  SCALAR_ID                       407334 non-null  int64  
 11  VECTOR                          407334 non-null  object 
 12  COORDINATE                      407334 non-null  object 
 13  VALUE                           405246 non-null  float64
 14  STATUS                          2088 non-null    object 
 15  SYMBOL                          0 non-null       float64
 16  TERMINATED                      0 non-null       float64
 17  DECIMALS                        407334 non-null  int64  
dtypes: float64(3), int64(4), object(11)
memory usage: 55.9+ MB
```

In [5]:
```
print('- Duplicates in Canadian data :',df_ca.duplicated().sum()) # Check for duplicates
print('- List of Years: ',df_ca['REF_DATE'].unique()) # Explore how many years
```

```
- Duplicates in Canadian data : 0
- List of Years:  [2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
 2014 2015 2016 2017 2018 2019 2020 2021 2022]
```

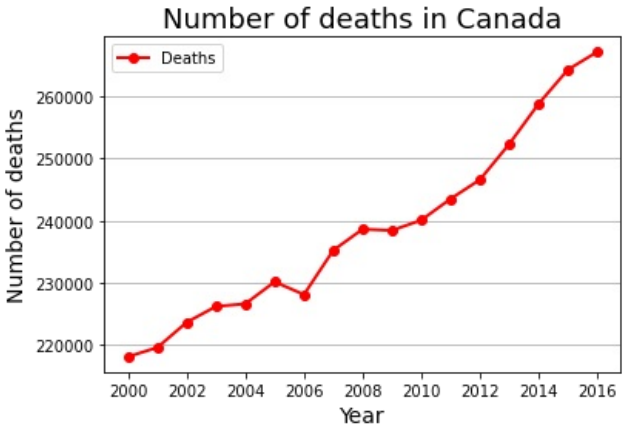
In [ ]:
```
# Explore all the causes --- Some EDA steps will not be displayed due to Page Limit
#df_ca['Leading causes of death (ICD-10)'].unique()

# Explore Age segmentation --- Some EDA steps will not be displayed due to Page Limit
#df_ca['Age at time of death'].unique()
```

Data cleaning

In [6]:
```
df_ca1 = df_ca[df_ca['Leading causes of death (ICD-10)'] != 'Total, all causes of death [A00-Y89]'] # Filter: Not
df_ca1 = df_ca1[df_ca1['Age at time of death'] == 'Age at time of death, all ages'] # Filter Age
df_ca1 = df_ca1[df_ca1['Characteristics'].str.contains("Number of deaths")] # Filter Characteristics
df_ca1 = df_ca1[df_ca1['Sex'].str.contains("Both sexes")] # Filter Both sexes
df_ca1 = df_ca1[(df_ca1['REF_DATE'] >= 2000) & (df_ca1['REF_DATE'] <= 2016)] # Define the period to align with US
```

Are Canadians facing increasing, decreasing, or steady likelihood of death?

In [7]:
```
df_total1 = df_ca1.groupby(['REF_DATE'])['VALUE'].sum().reset_index()
df_total1['VALUE'] = df_total1['VALUE'].astype(int)
data = {'Year' : df_total1['REF_DATE'],'Deaths' : df_total1['VALUE']}
dfp = pd.DataFrame(data)
dfp.plot(x='Year', y='Deaths', marker='o', color='r', linewidth=2, kind='line') # Plotting a line chart
plt.title('Number of deaths in Canada', fontsize=18) # Customize steps
plt.xlabel('Year', fontsize=14)
plt.ylabel('Number of deaths', fontsize=14)
plt.ticklabel_format(scilimits=(-5, 8))
plt.grid(axis='y')
plt.show()
```

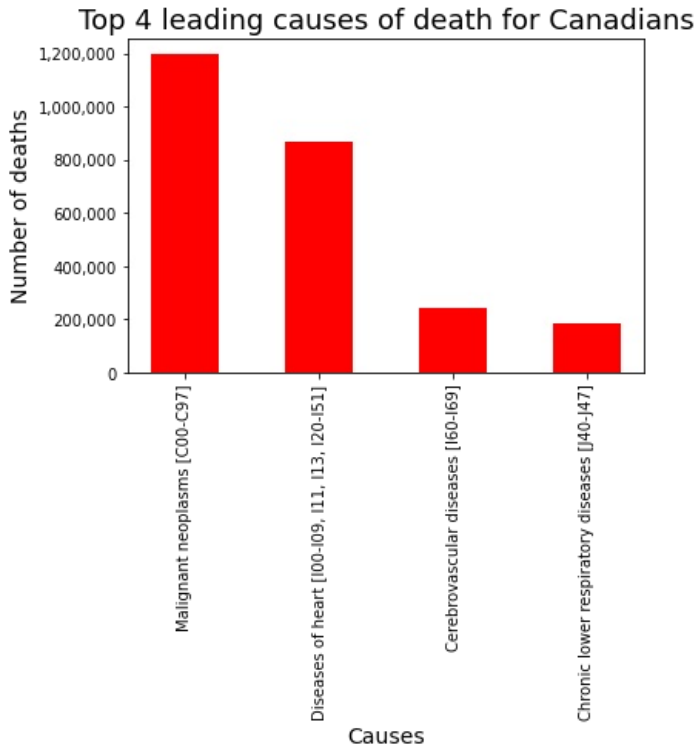


- Analysis: There is an increasing trend of number of deaths throughout the year in Canada, similar with the upward trend in the US.

What are the 4 leading causes of death for Canadians?

In [8]:
```
df_ca2 = df_ca1[df_ca1['Leading causes of death (ICD-10)'] != 'Other causes of death']
df_ca2 = df_ca2.groupby(['Leading causes of death (ICD-10)'])['VALUE'].sum()
sorted_grouped = df_ca2.sort_values(ascending=False)
df_plot2 = sorted_grouped.nlargest(4).reset_index()
plt.figure()
cause = df_plot2['Leading causes of death (ICD-10)']
death = df_plot2['VALUE']
plt.bar(cause, death, color = 'r', width= 0.5) # Plotting a bar chart
plt.xticks(cause, rotation = 90) # Customize steps
```

```
plt.xlabel('Causes', fontsize=14)
plt.ylabel('Number of deaths', fontsize=14)
plt.title('Top 4 leading causes of death for Canadians', fontsize=18)
def format_func(value, tick_number):
    return f'{int(value):,}'
plt.gca().yaxis.set_major_formatter(FuncFormatter(format_func))
plt.show()
```



Analysis:

- "Malignant neoplasms" commonly referred to as CANCER (source:https://my.clevelandclinic.org/health/diseases/22319-malignant-neoplasm);
- "Cerebrovascular diseases" commonly referred to as STROKE (source:https://my.clevelandclinic.org/health/diseases/24205-cerebrovascular-disease);
- "Chronic lower respiratory diseases" (CLRD).

Among of all diseases, it can be seen that the top 4 causes of deaths are CANCER, Diseases of heart, STROKE and CLRD, quite similar with the leading causes in the US.

Are there year-by-year changes in the 4 leading causes of death in Canada?

In [9]:

```
df_ca3 = df_ca1[df_ca1['Leading causes of death (ICD-10)'] != 'Other causes of death']
rows = []
for yr in df_ca3['REF_DATE'].unique():
    ca_cause_yrdf = df_ca3[df_ca3['REF_DATE'] == yr].groupby(['Leading causes of death (ICD-10)'])['VALUE'].sum()
    ca_cause_yrdf.sort_values(ascending=False, inplace=True)
    cause_list = list(ca_cause_yrdf.keys())[:4] + [None] * (4 - len(ca_cause_yrdf)) # Get the top 4 causes
    rows.append({'REF_DATE': yr, '1st cause': cause_list[0], '2nd cause': cause_list[1],
                 '3rd cause': cause_list[2], '4th cause': cause_list[3]})  # Append the row to the list
df_ca4 = pd.DataFrame(rows)
df_ca4.sort_values(by='REF_DATE', ascending=True, inplace=True)
display(df_ca4)
```

	REF_DATE	1st cause	2nd cause	3rd cause	4th cause
0	2000	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
1	2001	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
2	2002	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
3	2003	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
4	2004	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]

5	2005	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
6	2006	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
7	2007	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
8	2008	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
9	2009	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
10	2010	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Accidents (unintentional injuries) [V01-X59, Y...
11	2011	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
12	2012	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Accidents (unintentional injuries) [V01-X59, Y...
13	2013	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
14	2014	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
15	2015	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Chronic lower respiratory diseases [J40-J47]
16	2016	Malignant neoplasms [C00-C97]	Diseases of heart [I00-I09, I11, I13, I20-I51]	Cerebrovascular diseases [I60-I69]	Accidents (unintentional injuries) [V01-X59, Y...

- Analysis: In terms of ranking in Canada, CANCER, Diseases of heart and STROKE remain stably top 1, 2 and 3 respectively throughout the years. While 4th place is held between CLRD and Accidents (Unintentional Injuries). Meanwhile in the US, Heart disease and CANCER are consistently top 1 and 2 , respectively, with the 3rd and 4th are normally STROKE, CLRD and Unintentional Injuries.

Conclusion

In summary, given the assumptions stated in the beginning, along with questions got answered, it could be confirmed that the data of causes of death between Canada and the US is quite similar with regards to trend and top causes. Further investigations may examine the role of sociology factors (e.g. Sexes, Income level, Geography,...) to tackle the trends and whether viable solutions that has proven successful in one country could be applied to the other country.