Assignment 4

Before working on this assignment please read these instructions fully. In the submission area, you will notice that you can click the link to **Preview the Grading** for each step of the assignment. This is the criteria that will be used for peer grading. Please familiarize yourself with the criteria before beginning the assignment.

This assignment requires that you to find **at least** two datasets on the web which are related, and that you visualize these datasets to answer a question with the broad topic of **religious events or traditions** (see below) for the region of **Gardena, California, United States**, or **United States** more broadly.

You can merge these datasets with data from different regions if you like! For instance, you might want to compare **Gardena, California, United States** to Ann Arbor, USA. In that case at least one source file must be about **Gardena, California, United States**.

You are welcome to choose datasets at your discretion, but keep in mind **they will be shared with your peers**, so choose appropriate datasets. Sensitive, confidential, illicit, and proprietary materials are not good choices for datasets for this assignment. You are welcome to upload datasets of your own as well, and link to them using a third party repository such as github, bitbucket, pastebin, etc. Please be aware of the Coursera terms of service with respect to intellectual property.

Also, you are welcome to preserve data in its original language, but for the purposes of grading you should provide english translations. You are welcome to provide multiple visuals in different languages if you would like!

As this assignment is for the whole course, you must incorporate principles discussed in the first week, such as having as high data-ink ratio (Tufte) and aligning with Cairo's principles of truth, beauty, function, and insight.

Here are the assignment instructions:

- State the region and the domain category that your data sets are about (e.g., **Gardena, California, United States** and **religious events or traditions**).
- You must state a question about the domain category and region that you identified as being interesting.
- You must provide at least two links to available datasets. These could be links to files such as CSV or Excel files, or links to websites

which might have data in tabular form, such as Wikipedia pages.
- You must upload an image which addresses the research question you stated. In addition to addressing the question, this visual should follow Cairo's principles of truthfulness, functionality, beauty, and insightfulness.
- You must contribute a short (1-2 paragraph) written justification of how your visualization addresses your stated research question.

What do we mean by **religious events or traditions**? For this category you might consider calendar events, demographic data about religion in the region and neighboring regions, participation in religious events, or how religious events relate to political events, social movements, or historical events.

Tips

- Wikipedia is an excellent source of data, and I strongly encourage you to explore it for new data sources.
- Many governments run open data initiatives at the city, region, and country levels, and these are wonderful resources for localized data sources.
- Several international agencies, such as the United Nations (http://data.un.org/), the World Bank (http://data.worldbank.org/), the Global Open Data Index (http://index.okfn.org/place/) are other great places to look for data.
- This assignment requires you to convert and clean datafiles. Check out the discussion forums for tips on how to do this from various sources, and share your successes with your fellow students!

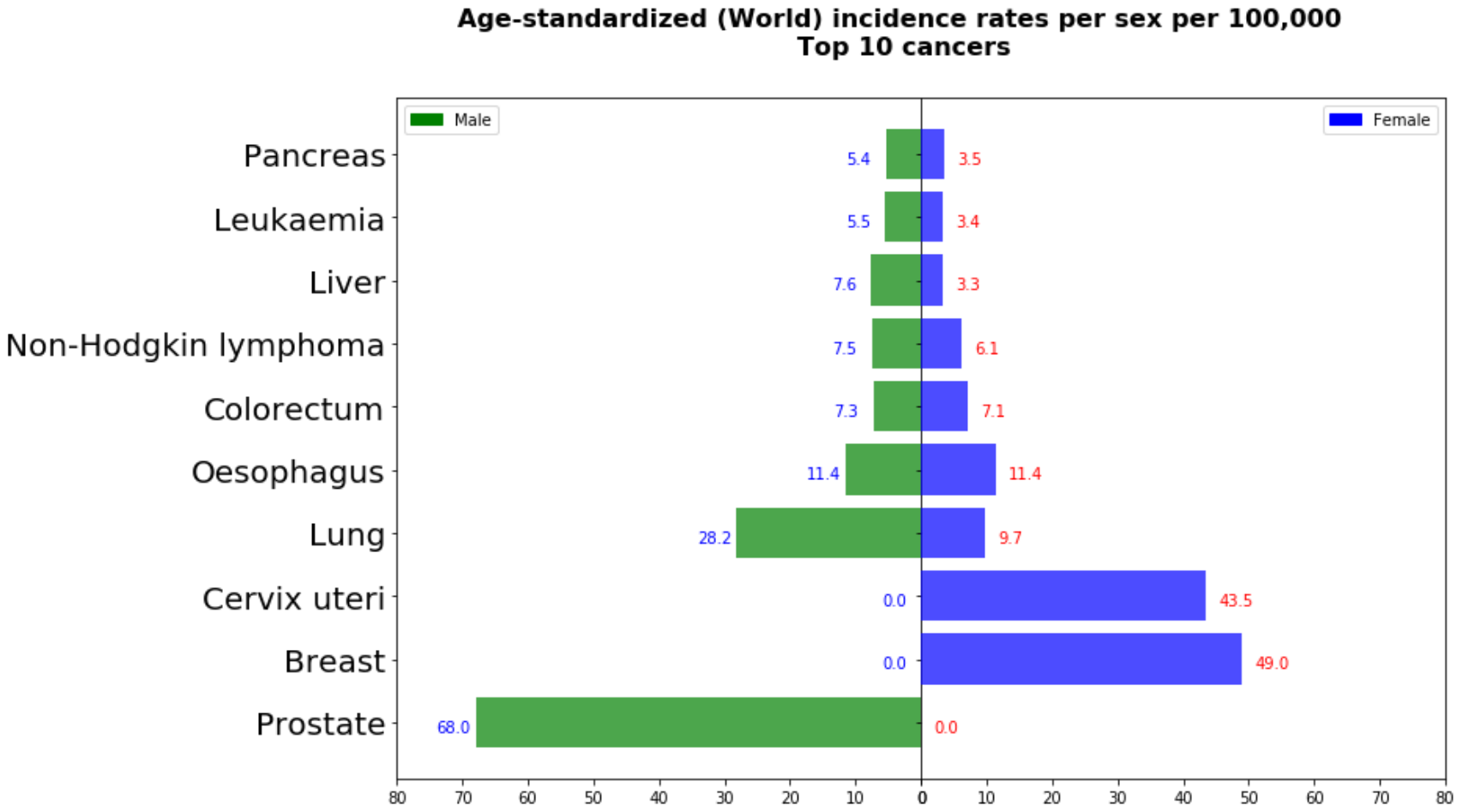
Example

Looking for an example? Here's what our course assistant put together for the **Ann Arbor, MI, USA** area using **sports and athletics** as the topic. Example Solution File (./readonly/Assignment4_example.pdf)

In [3]:
```
#Data for plot

cancer_male={'Prostate':68,'Breast':0,'Cervix uteri':0,'Lung':28.2,'Oesophagus':11.4,'Colorectum'
:7.3,'Non-Hodgkin lymphoma':7.5,'Liver':7.6,'Leukaemia':5.5,'Pancreas':5.4}
cancer_female={'Prostate':0,'Breast':49,'Cervix uteri':43.5,'Lung':9.7,'Oesophagus':11.4,'Colorec
tum':7.1,'Non-Hodgkin lymphoma':6.1,'Liver':3.3,'Leukaemia':3.4,'Pancreas':3.5}

cancer_rate_male = list(cancer_male.values())
cancer_type = list(cancer_male.keys())
cancer_rate_female = list(cancer_female.values())
```

In [4]:
```
import matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec
import numpy as np
from matplotlib.ticker import MaxNLocator
import matplotlib.patches as mpatches

#adding two subplots and plot title
fig = plt.figure(figsize=(12,8))
fig.suptitle('Age-standardized (World) incidence rates per sex per 100,000 \nTop 10 cancers', fon
tsize=16,fontweight='bold')

ax1 = fig.add_subplot(1,2,1, aspect = "auto")
ax2 = fig.add_subplot(1,2,2, aspect = "auto", sharey = ax1)

# to delete the gap between the subplots and make second subplot y axis ticks invisible
plt.subplots_adjust(wspace=0, hspace=0)
plt.setp(ax2.get_yticklabels(), visible=False)

#x and y axis data for barh plots
y_pos = np.arange(len(cancer_type))
cancer_rate_male=np.array(cancer_rate_male)
cancer_rate_female=np.array(cancer_rate_female)
```

```
#first subpolot, ytick label, x tick limits and pasting the bar value next to bars
rect1=ax1.barh(y_pos, cancer_rate_male, color='green',zorder=10,alpha=0.7)
ax1.set(title='')
ax1.set(yticks=y_pos, yticklabels=cancer_type)
ax1.set_xlim(xmin=0.0, xmax=80)
ax1.invert_xaxis()
for i, v in enumerate(cancer_rate_male):
    ax1.text(v + 6, i -0.15, str(v), color='blue')

#second subpolot, ytick label, x tick limits and pasting the bar value next to bars
rect2 =ax2.barh(y_pos, cancer_rate_female, color='blue',zorder=10,alpha=0.7)
ax2.set(title='')
ax2.set_xlim(xmin=0.0, xmax=80)
ax1.yaxis.set_tick_params(labelsize=20)
for i, v in enumerate(cancer_rate_female):
    ax2.text(v + 2, i -0.15, str(v), color='red')

#subplots legend
M = mpatches.Patch(color='green', label='Male')
F = mpatches.Patch(color='blue', label='Female')
ax1.legend(handles=[M], loc=2)
ax2.legend(handles=[F], loc=1)


plt.show()
```



In [10]:
```
import pandas as pd
from scipy import stats
#read cancer rate data from external website for Top 50 countries
cancer = pd.read_excel(r'cancer_in_all.xlsx')

#read GDP per Capita from 2010 to 2015 from world bank data
GDP_pc = pd.read_excel(r'gdp_pc.xlsx')
GDP_pc = GDP_pc.replace("np.nan","0")
```

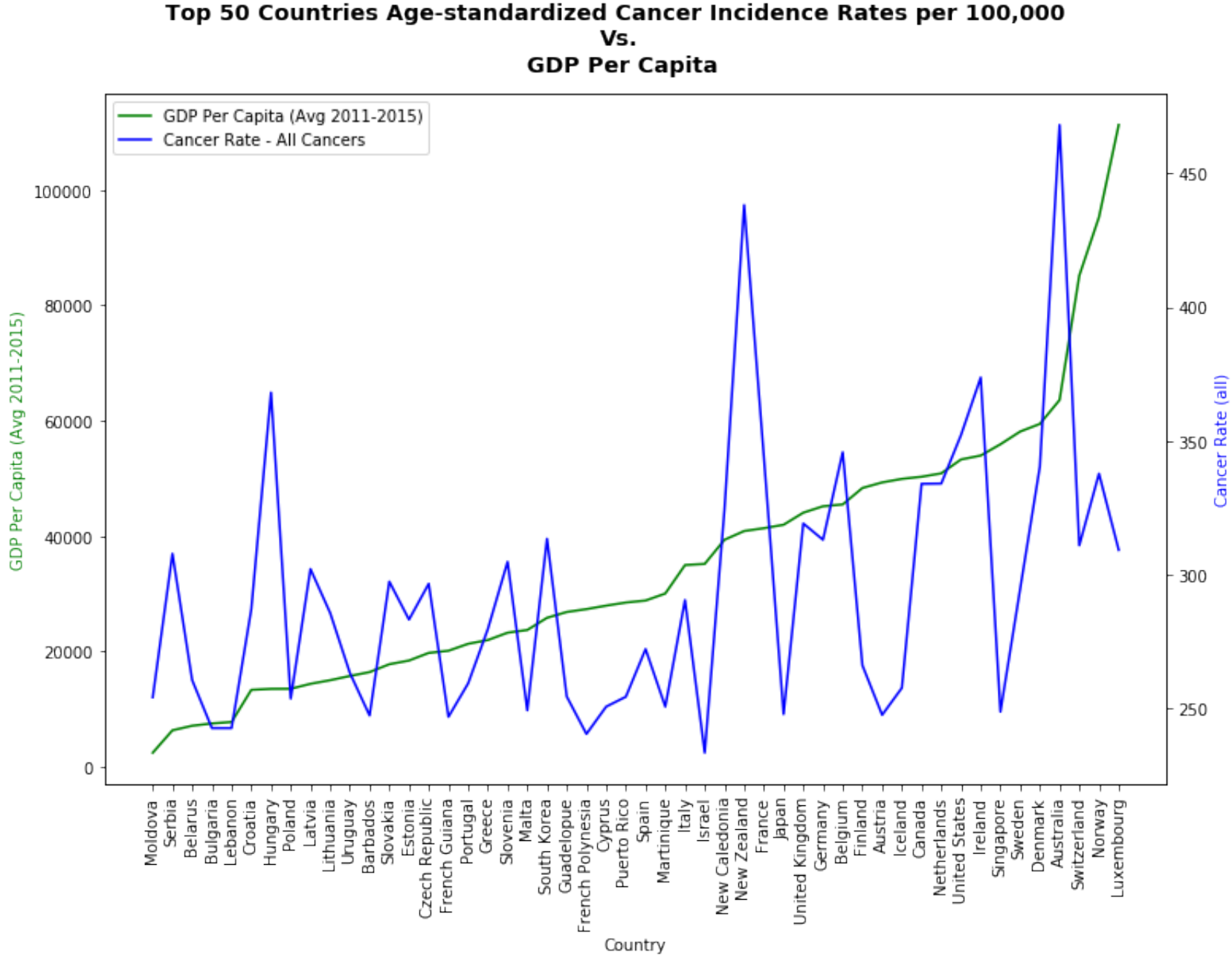
```
# select Top 50 cancer countries
cont=cancer['Country']
cancer_g=pd.DataFrame()

# build dataframe for Top 50 country avg GDP per Capita, Cancer Rate
for c in cont:
    en_cancer=GDP_pc[GDP_pc['Country'] == c]
    en_cancer=en_cancer.dropna()
    en_cancer=en_cancer.set_index(['Country'])
    cancer_g=cancer_g.append(en_cancer,ignore_index=False)
cancer_g = cancer_g.replace("..",np.NaN)
cancer_g['2010'] = cancer_g['2010'].str.replace(',', '')
cancer_g['2011'] = cancer_g['2011'].str.replace(',', '')
cancer_g['2012'] = cancer_g['2012'].str.replace(',', '')
cancer_g['2013'] = cancer_g['2013'].str.replace(',', '')
cancer_g['2014'] = cancer_g['2014'].str.replace(',', '')
cancer_g['2015'] = cancer_g['2015'].str.replace(',', '')

cancer_g['2010'] = cancer_g['2010'].astype(float)
cancer_g['2011'] = cancer_g['2011'].astype(float)
cancer_g['2012'] = cancer_g['2012'].astype(float)
cancer_g['2013'] = cancer_g['2013'].astype(float)
cancer_g['2014'] = cancer_g['2014'].astype(float)
cancer_g['2015'] = cancer_g['2015'].astype(float)
cancer_g.drop(['2010'],axis=1,inplace=True)

cancer_g.replace(0.0, np.NaN, inplace=True)
cancer_g['avg']= cancer_g.mean(axis=1, skipna=True)
cancer_g = cancer_g.reset_index()


cancer_merge = pd.merge(cancer_g,cancer,left_on=['Country'],right_on=['Country'],left_index=False
, right_index=False,how ='inner')
cancer_merge_f=cancer_merge[['Country','avg','cancer rate']]
cancer_merge_f.rename(columns = {'avg':'GDP_AVG'}, inplace = True)
cancer_merge_f.sort_values(by=['GDP_AVG'], inplace=True)
cancer_merge_f=cancer_merge_f.set_index(['Country'])
```

```
# perform t stat for countries below median GDP per Capita and Countries above median GDP per Cap
ita
median= cancer_merge_f['GDP_AVG'].median()

cancer_rate_b=cancer_merge_f['cancer rate'].where(cancer_merge_f['GDP_AVG'] <median).dropna()
cancer_rate_a=cancer_merge_f['cancer rate'].where(cancer_merge_f['GDP_AVG'] >median).dropna()
mean_a = cancer_rate_b.mean(axis=0)
mean_b = cancer_rate_a.mean(axis=0)


s, p = stats.ttest_ind(cancer_rate_b, cancer_rate_a)
print(s,p,mean_a,mean_b)

#Plot line plot for GDP Per Capita and Cancer Rate
fig, ax1 = plt.subplots(figsize=(12,8))
fig.suptitle('Top 50 Countries Age-standardized Cancer Incidence Rates per 100,000 \nVs.\n GDP Pe
r Capita', fontsize=14,fontweight='bold')
x=range(len(cancer_merge_f))
ax2 = ax1.twinx()
gdp_avg = ax1.plot(x, cancer_merge_f['GDP_AVG'], 'g-',label='GDP Per Capita (Avg 2011-2015)')
cancer_rate = ax2.plot(x, cancer_merge_f['cancer rate'],'b-',label='Cancer Rate - All Cancers')

ax1.set_xlabel('Country')
ax1.set_ylabel('GDP Per Capita (Avg 2011-2015)', color='g')
ax2.set_ylabel('Cancer Rate (all)', color='b')
ax1.set(xticks=x, xticklabels=cancer_merge_f.index)

for tick in ax1.get_xticklabels():
    tick.set_rotation(90)
ax1.xaxis.set_tick_params(labelsize=10)

lns = gdp_avg+cancer_rate
labs = [l.get_label() for l in lns]
ax1.legend(lns, labs, loc=0)

#CalCulate Correlation between GDP Per Capita and Cancer Rate
```

```
corr_all = cancer_merge_f['GDP_AVG'].corr(cancer_merge_f['cancer rate'])

plt.show()
```

```
/opt/conda/lib/python3.6/site-packages/pandas/core/frame.py:2834: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
#indexing-view-versus-copy
  **kwargs)
/opt/conda/lib/python3.6/site-packages/ipykernel/__main__.py:44: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
#indexing-view-versus-copy
```

```
-3.08417530511 0.00338088429915 273.944 314.224
```


Top 50 Countries Age-standardized Cancer Incidence Rates per 100,000
Vs.
GDP Per Capita
GDP Per Capita (Avg 2011-2015)
Cancer Rate - All Cancers
GDP Per Capita (Avg 2011-2015)
Cancer Rate (all)
Country
0
20000
40000
60000
80000
100000
250
300
350
400
450
Moldova
Serbia
Belarus
Bulgaria
Lebanon
Croatia
Hungary
Poland
Latvia
Lithuania
Uruguay
Barbados
Slovakia
Estonia
Czech Republic
French Guiana
Portugal
Greece
Slovenia
Malta
South Korea
Guadelopue
French Polynesia
Cyprus
Puerto Rico
Spain
Martinique
Italy
Israel
New Caledonia
New Zealand
France
Japan
United Kingdom
Germany
Belgium
Finland
Austria
Iceland
Canada
Netherlands
United States
Ireland
Singapore
Sweden
Denmark
Australia
Switzerland
Norway
Luxembourg

In [ ]: