NYPD Final Project: Cancer Atlas Data Analysis

Kathy (Kasia) Wiech

February 9, 2022

Abstract

This report discusses the code I wrote to analyze the Cancer Atlas 3er (available at www.canceratlas.cancer.org.

1 Introduction

In 56 countries, cancer is the number one cause of death for those under 70 years old (See Figure 1). Various factors play into the development of cancer, including lifestyle and genetics. Reports like the Cancer Atlas provide value information on what factors can be used to understand what risk factors are associated with higher cancer incidences across the globe.

2 Data Analysis

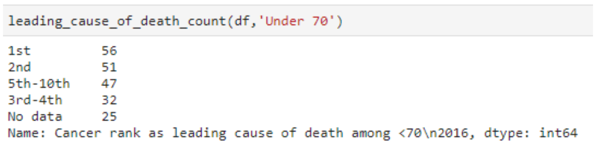
2.1 Warning Labels Analysis

I was interested in exploring whether graphic warning labels on cigarette packaging had any impact on lung cancer rates. The Cancer Atlas has information about the lung cancer incidence for males and females in countries across the world, in addition to information about the kind of packaging included on cigarette cartridges. I used the built in pandas mean method and the matplotlib histogram method to create functions that returned the mean lung cancer incidence for both genders across countries with different types of graphic packaging. The function, $warning\_labels\_analysis(dataframe)$ is available through the $warning\_labels.py$ module.

The mean of countries with graphic packaging labels (for example, with pictures of cancer) actually tend to have higher cancer incidence for both men and women than countries that do not.See figure 2 for the mean lung cancer incidence for men. This is likely because graphic packing was introduce after high lung cancer incidence was recognized as a problem and lung cancer tends to be a slow developing disease. It would be interesting to explore how lung cancer incidence changes over time after the introduction of graphic packaging.

2.2 Smoking Proportion Analysis

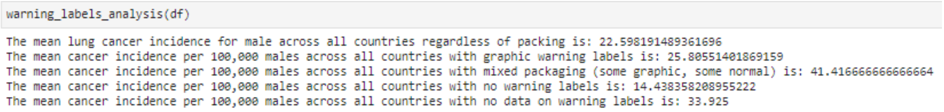
Next, I wanted to look at a well know risk factor for lung cancer: smoking history. The Cancer Atlas provides information about the proportion of men and women who smoke daily in countries across the world. I used pandas built in correlation and scatter plot methods to create a function that explores how the proportion daily smoking correlated with lung cancer incidence across the world. Figure 3



```
leading_cause_of_death_count(df,'Under 70')

1st          56
2nd          51
5th-10th     47
3rd-4th      32
No data      25
Name: Cancer rank as leading cause of death among <70\n2016, dtype: int64
```

Figure 1: This screenshot shows the value counts of the column "Cancer rank as leading cause of death among under 70".



```
warning_labels_analysis(df)

The mean lung cancer incidence for male across all countries regardless of packing is: 22.598191489361696
The mean cancer incidence per 100,000 males across all countries with graphic warning labels is: 25.80551401869159
The mean cancer incidence per 100,000 males across all countries with mixed packaging (some graphic, some normal) is: 41.416666666666664
The mean cancer incidence per 100,000 males across all countries with no warning labels is: 14.438358208955222
The mean cancer incidence per 100,000 males across all countries with no data on warning labels is: 33.925
```

Figure 2: This screenshot shows the first part of the output of the $warning\_labels\_analysis(dataframe)$ As seen, countries with graphic warning labels on average have higher lung cancer incidence for men.

shows the scatter plots and correlation coefficients for males and females who are daily smokers. The function *smoking_cancer_analysis* is available through the *smoking_analysis.py* module.

In the data for men, there is an outlier (Indonesia, which reports that 66.6% of men are daily smokers but only 19.4 lung cancer cases per 100,000 males). Even so, there is still a relatively strong correlation between the proportion of daily smokers and lung cancer incidence.

Within the *smoking_analysis.py* module, there are functions that allow one to output a dataframe with the countries with highest/lowest proportion of smokers (*smoke_rank*), the highest/lowest lung cancer incidence (*cancer_rank*) and then a dataframe of countries that appear on both the highest/lowest proportion of smokers list and the highest/lowest incidence of lung cancer list (*compare_cancer_smoke*). Each of the aforementioned functions accepts three arguments: a dataframe, a string that decides which ranking will be used ('Most' for highest proportion of cancer/smokers, 'Least' for the lowest proportion of cancer/smokers, and 'Both' to show both), and an integer to decide how many countries to include in the dataframe. Figure 4 shows the output of *compare_cancer_smoke* for countries that appear on both the top 20 lists for lung cancer and smokers.

2.3 Air Pollution Analysis

I also decided to see if air pollution was related to lung cancer in any way. Using similar functions to the ones present in the module *smoking_analysis.py* I was able to look at the correlation coefficient and scatter plots for indoor and outdoor air pollution and lung cancer incidence. I was surprised to find that both outdoor and indoor air pollution were correlated with lower lung cancer rates. This could be because countries with more air pollution tend to be less developed and are therefore less likely to have good reporting methods for various cancers. Figure 5 shows the output for *outdoor_pol_analysis*, which is available from the *pollution_analysis.py* module. Functions included in the *pollution_analysis.py* module are *outdoor_pol_cancer_analysis*, *indoor_pol_cancer_analysis*, *outdoor_pol_rank* and *indoor_pol_rank*.

2.4 Radiotherapy

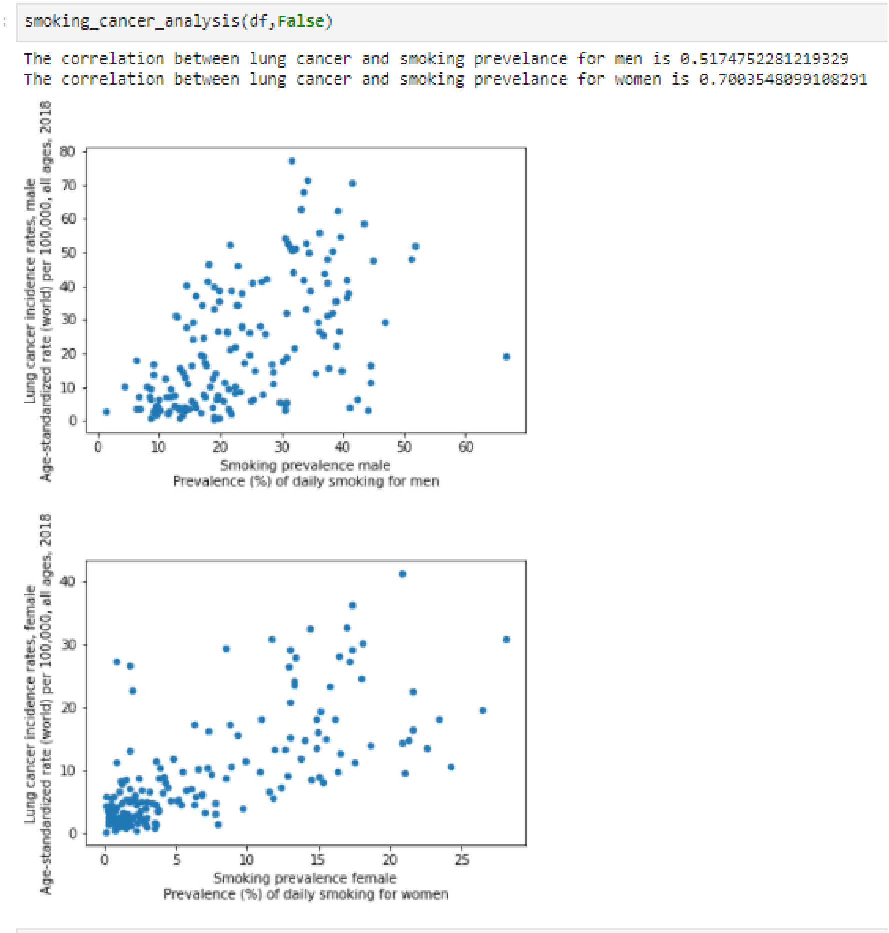
The module *radio_common.py* has the function *radiotherapy_cancer_survivor_analysis* which returns the correlation coefficient and a scatter plot for the availability of radiotherapy machines and the number of cancer survivors in a country. Radiotherapy availability is positively correlated with the number of cancer survivors in a country. Figure 6 shows the output of the *radiotherapy_cancer_survivor_analysis* function.

2.5 Common Cancers and Cause of Death

Also in the $radio_common.py$ module are two additional functions, $leading\_cause\_of\_death\_count$ and $top_cancers$. The output of $leading\_cause\_of\_death_count$ outputs the value counts of the "Cancer rank as a leading cause of death among age group" and is pictured in Figure 1. There are two possible age groups, 30-69 and under 70. The $top_cancers$ function returns the most common cancer and most common cause of cancer death for both males in females and its output is pictured in Figure 7.

2.6 Future Work

In the future, I think it would be interesting to explore how the Human Development Index affects reported cancer cases and cancer survivors. Further more, I think it would be interesting to see how the proportion of smokers relates to air pollution and the Human Development Index.

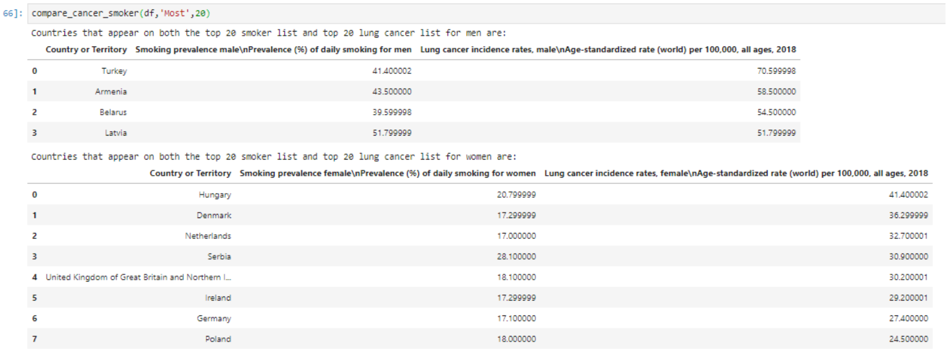


```
smoking_cancer_analysis(df,False)
```

```
The correlation between lung cancer and smoking prevelance for men is 0.5174752281219329
The correlation between lung cancer and smoking prevelance for women is 0.7003548099108291
```

Lung cancer incidence rates, male
Age-standardized rate (world) per 100,000, all ages, 2018
0
10
20
30
40
50
60
70
80
0
10
20
30
40
50
60
Smoking prevalence male
Prevalence (%) of daily smoking for men
Lung cancer incidence rates, female
Age-standardized rate (world) per 100,000, all ages, 2018
0
10
20
30
40
0
5
10
15
20
25
Smoking prevalence female
Prevalence (%) of daily smoking for women

Figure 3: The output of the function *smoking_cancer_analysis*$(dataframe, outlier)$ According to the Cancer Atlas, daily smoking is more heavily correlated with lung cancer for women than it is for me.



```
66]: compare_cancer_smoker(df,'Most',20)
```

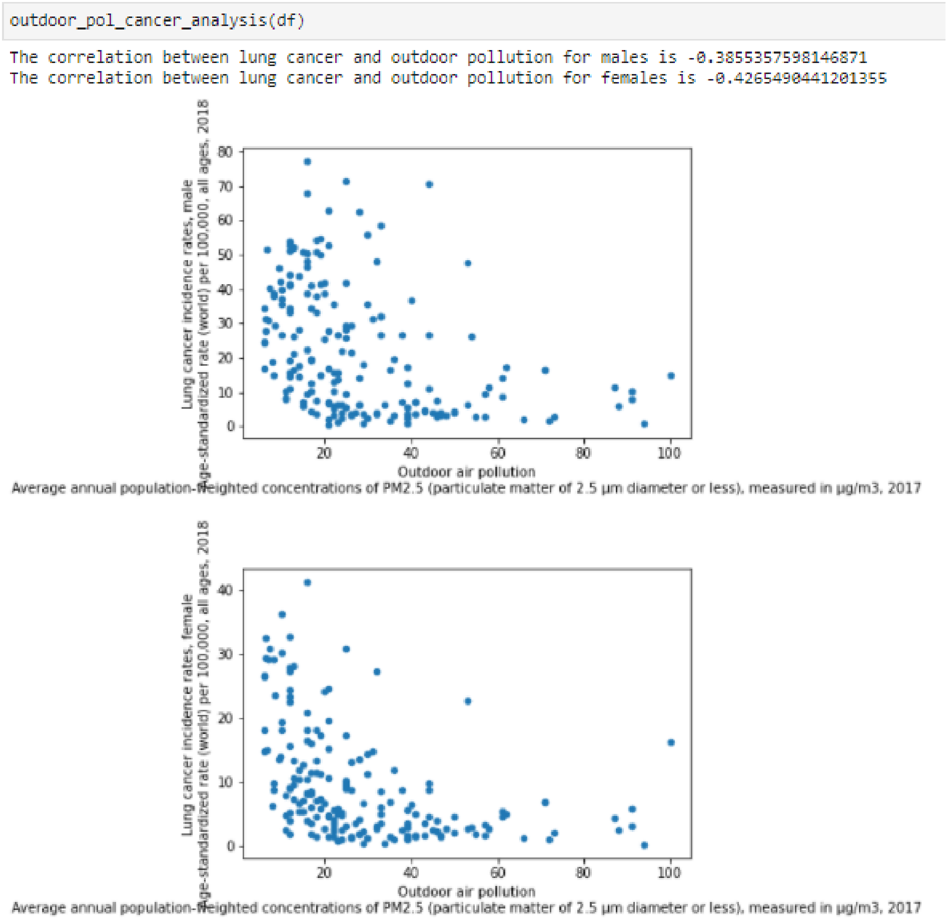
```
Countries that appear on both the top 20 smoker list and top 20 lung cancer list for men are:
```

	Country or Territory	Smoking prevalence male\nPrevalence (%) of daily smoking for men	Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018
0	Turkey	41.400002	70.599998
1	Armenia	43.500000	58.500000
2	Belarus	39.599998	54.500000
3	Latvia	51.799999	51.799999

```
Countries that appear on both the top 20 smoker list and top 20 lung cancer list for women are:
```

	Country or Territory	Smoking prevalence female\nPrevalence (%) of daily smoking for women	Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018
0	Hungary	20.799999	41.400002
1	Denmark	17.299999	36.299999
2	Netherlands	17.000000	32.700001
3	Serbia	28.100000	30.900000
4	United Kingdom of Great Britain and Northern I...	18.100000	30.200001
5	Ireland	17.299999	29.200001
6	Germany	17.100000	27.400000
7	Poland	18.000000	24.500000

Figure 4: The output of *compare_cancer_smoke* for countries on both the top 20 list for smokers and lung cancer.

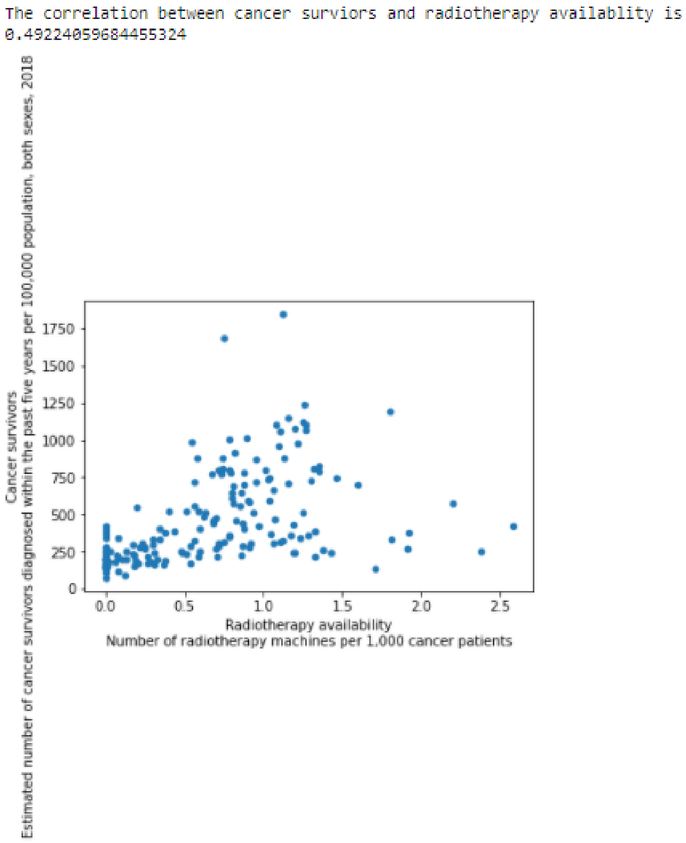


```
outdoor_pol_cancer_analysis(df)
```

```
The correlation between lung cancer and outdoor pollution for males is -0.3855357598146871
The correlation between lung cancer and outdoor pollution for females is -0.4265490441201355
```

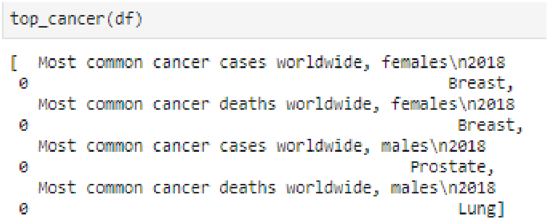
Lung cancer incidence rates, male
Age-standardized rate (world) per 100,000, all ages, 2018
80
70
60
50
40
30
20
10
0
20
40
60
80
100
Outdoor air pollution
Average annual population-weighted concentrations of PM2.5 (particulate matter of 2.5 µm diameter or less), measured in µg/m3, 2017
Lung cancer incidence rates, female
Age-standardized rate (world) per 100,000, all ages, 2018
40
30
20
10
0
20
40
60
80
100
Outdoor air pollution
Average annual population-weighted concentrations of PM2.5 (particulate matter of 2.5 µm diameter or less), measured in µg/m3, 2017

Figure 5: This figure shows the output of the *outdoor_pol_cancer_analysis* function. Outdoor air pollution is negatively correlated with lung cancer incidence. This means countries with less air pollution are actually more likely to have a higher incidence of lung cancer cases in both male and females.



```
The correlation between cancer surviors and radiotherapy availablity is
0.49224059684455324
```

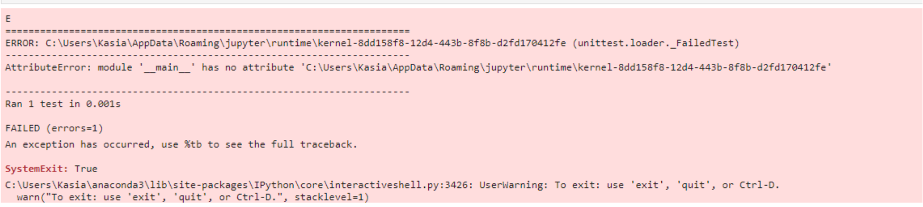
Figure 6: This figure displays the output of the *radiotherapy_cancer_survivor_analysis* function. Radiotherapy availability is positively correlated with the number of cancer survivors in a country.



```
top_cancer(df)
```

```
[  Most common cancer cases worldwide, females\n2018
 0                                             Breast,
   Most common cancer deaths worldwide, females\n2018
 0                                              Breast,
   Most common cancer cases worldwide, males\n2018
 0                                         Prostate,
   Most common cancer deaths worldwide, males\n2018
 0                                              Lung]
```

Figure 7: The output of the *top_cancer* function. The most common cancer and most common cancer deaths for women across the globe is breast cancer, while the most common cancer for men worldwide is prostate cancer and the most common cancer death is lung cancer.



```
E
======================================================================
ERROR: C:\Users\Kasia\AppData\Roaming\jupyter\runtime\kernel-8dd158f8-12d4-443b-8f8b-d2fd170412fe (unittest.loader._FailedTest)
----------------------------------------------------------------------
AttributeError: module '__main__' has no attribute 'C:\Users\Kasia\AppData\Roaming\jupyter\runtime\kernel-8dd158f8-12d4-443b-8f8b-d2fd170412fe'

----------------------------------------------------------------------
Ran 1 test in 0.001s

FAILED (errors=1)
An exception has occurred, use %tb to see the full traceback.

SystemExit: True
C:\Users\Kasia\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3426: UserWarning: To exit: use 'exit', 'quit', or Ctrl-D.
  warn("To exit: use 'exit', 'quit', or Ctrl-D.", stacklevel=1)
```

Figure 8: Unfortunately, my unittest failed.

3 Unittest and Profiling

3.1 Unittest

The unittest code is found in the $df_testing.py$ module. I tried to test the opening of the excel file and the dataframe, but it did not go well for me. The results are in Figure 10.

3.2 Profiling

I also tried Profiling my code, using *%lprun*, and *memit*. The output of the Unfortunately, some of these profiling methods would run on all of my functions, and this is something I could improve. Most of my functions were quick, with a runtime of 1-e07 seconds. I believe I could improve many of these functions if I could find an effective way to remove "No data" rows all at once instead of column by column.