George Washington University Cancer Center

Cancer Data Visualizer

Technical Documentation

*Last updated*: 3/23/20

CONTENTS

* [Purpose](#bookmark=id.rdkhtgfselo3)
* [Regions](#bookmark=id.tajdnah1js8l)
* [Data Sources](#bookmark=id.xjeadvd0hy4n)
  + [Cancer Rates - DC, Maryland & Virginia](#bookmark=id.z28mvb9t6deu)
  + [Cancer Rates - DC wards](#bookmark=id.2p5yqd1rkrm1)
  + [Risk & Protective Factors - American Community Survey - Maryland & Virginia](#bookmark=id.vy7zzlofq1us)
  + [Risk & Protective Factors - American Community Survey - DC & DC wards](#bookmark=id.u4zyc6azc8hz)
  + [Risk & Protective Factors - Health & Risk Behaviors](#bookmark=id.spznsgy0kxwj)
  + [Risk & Protective Factors - Other Variables](#bookmark=id.928r7nr5tgm9)
* [Pre-Processing](#bookmark=id.5f0sx97xx4t)
  + [Maps](#bookmark=id.otr4agms8ahg)
  + [Pre-Processing - Cancer Data](#bookmark=id.u2jw54ksh2ag)
  + [Pre-Processing - Risk & Protective Factors](#bookmark=id.xfsvxboij1qj)
* [Shiny Architecture](#bookmark=id.8gvq28p3izwj)
  + [DEV Environment](#bookmark=id.pxj3twk7y9wh)
  + [TEST & PROD Environments](#bookmark=id.g7akh8vebgt0)
  + [Code Sharing](#bookmark=id.a3iz9javp9uj)

Purpose

This is meant to serve as a living document that describes how the GWCC Cancer Data Visualizer works internally. It describes all technologies in use and how they fit together. The main areas this document addresses include:

1. application architecture
2. how to reproduce any functionality or data
3. how to update the application to include new functionality or data

The main tool used in this project is the [R language](https://www.r-project.org/). It is used in two primary ways:

* pre-processing raw data into organized data files that are easily consumed by the visualizer
* coding the visualizer as a web app with the R package [Shiny](https://shiny.rstudio.com/) that allows end users to interact with the data

Regions

The GWCC Catchment Area spans the following areas:

* District of Columbia and its 8 wards
* 3 Maryland counties: Charles County, Montgomery County, Prince George’s County
* 4 Virginia counties: Arlington County, Fairfax County, Loudoun County, Prince William County
* 5 Virginia independent cities: Alexandria, Fairfax, Falls Church, Manassas, Manassas Park

Counties and independent cities are separate and both need to be included.

Certain data differ among regions when it comes to reporting standards or availability. The [Data Sources](#bookmark=id.xjeadvd0hy4n) section notes these discrepancies.

Data Sources

At a high level, all the data can be divided into two categories:

* cancer rates
* risk and protective factors

Wherever possible, data is obtained from APIs, but most of the time data files must be downloaded or requested.

*Cancer Rates - DC, Maryland & Virginia*

Cancer rates for DC, Maryland, and Virginia are taken from the Centers for Disease Control and Prevention. These data are available as text files from the following url: <https://www.cdc.gov/cancer/uscs/dataviz/download_data.htm>. When you download one of these zipped folders, only 2 files are required:

* BYAREA\_COUNTY - contains incidence and mortality rates by county and independent city
* BYAREA - contains incidence and mortality rates by state, including DC

The variable used from these files is *AGE\_ADJUSTED\_RATE*, which is the average annual age adjusted rate over 5 years. For example, the 2016 download link has rates over the time period 2012-2016.

If you look closely at this webpage, any of the downloads before 2015 do not have county-level files. Thus, only the 2011-2015 and 2012-2016 time periods are used for these data.

*Cancer Rates - DC wards*

DC ward-level cancer data was specially requested from the [DC Cancer Registry](https://dchealth.dc.gov/service/cancer-registry-0). These data are only available for the time period 2012-2016.

In addition, due to different reporting standards, ward cancer data is in general inflated compared to the rest of the GWCC Catchment Area.

*Risk & Protective Factors - American Community Survey - Maryland & Virginia*

The majority of county and independent city risk and protective factors comes from the American Community Survey’s [Census API](https://api.census.gov/data.html). This API is used to retrieve data from the *ACS 5-Year Data Profile* for each of the years 2013 to 2017. For example, the url <https://api.census.gov/data/2017/acs/acs5/profile?get=DP03_0128PE&for=county:013&in=state:51&key=6ab3340c16d514e7094fbdeff066406455377bea> returns the % below poverty level in Arlington County in 2017.

This is a list of variables currently used from the ACS:

* % Below Poverty Level
* Educational Attainment
* Employment Status
* Ethnicity
* Foreign-Born
* Health Insurance Coverage
* Housing Tenure
* Language Spoken at Home
* Median Age
* Median Household Income
* Population
* Race
* % Households with Rent Greater than 30 Percent of Household Income
* Vehicles per Housing Unit

*Risk & Protective Factors - American Community Survey - DC & DC wards*

The Census API does not contain data for DC wards. Instead, ACS data for DC and its wards are obtained from data files at this url: <https://planning.dc.gov/page/american-community-survey-acs-estimates>. These data files contain data for the same variables as in the Census API. The 5-Year files should be used here.

*Risk & Protective Factors - Health & Risk Behaviors*

The [Robert Wood Johnson Foundation](https://www.countyhealthrankings.org/app/) has data downloads available for just Maryland and Virginia counties. The following variables are used from this source:

* % of Children Eligible for Free Lunch
* % Diabetic
* % Diabetic Screening
* % Excessive Drinking
* HIV Prevalence
* Homicide Rate
* % Inadequate Social Support
* % Limited Access to Healthy Foods
* % Mammography Screening
* % Obesity
* % Physically Inactive
* % Poor or Fair Health
* Premature Mortality Rate
* % Single-Parent Households
* % Smoking
* Violent Crime Rate

In general, these data are unavailable for many regions and years. In addition, there is a time lag between the year of reporting and actual year.

*Risk & Protective Factors - Other Variables*

Air Quality Index is available for aggregate DC, Maryland counties, and Virginia counties via data files from the [Environmental Protection Agency](https://aqs.epa.gov/aqsweb/airdata/download_files.html).

Pre-Processing

The *Data* directory is organized as follows:

* Raw\_Data
  + American Community Survey - <https://planning.dc.gov/page/american-community-survey-acs-estimates>
    - 2009-2013 ACS 5-Year Esitmates-Ward\_0.xls
    - 2010-2014 ACS 5-Year Estimates-Ward\_0.xlsx
    - 2011-2015 Ward.xls
    - 2012-2016 ACS 5-Year Ward.xls
    - 2013-2017 ACS 5-Year Ward.xls
    - 2009-2013 ACS 5-Year Districtwide.xls
    - 2010-2014 ACS 5-Year Estimates-Districtwide\_0.xls
    - 2011-2015 Districtwide.xls
    - 2012-2016 ACS 5 -Year Districtwide.xls
    - 2013-2017 ACS 5-Year Districtwide.xls
  + DC Cancer Registry - specially requested
    - DC cases\_2012-2016\_All and top 4 by sex and ward\_Letterhead With CHA.csv
  + EPA - <https://aqs.epa.gov/aqsweb/airdata/download_files.html>
    - annual\_aqi\_by\_county\_2013.csv
    - annual\_aqi\_by\_county\_2014.csv
    - annual\_aqi\_by\_county\_2015.csv
    - annual\_aqi\_by\_county\_2016.csv
    - annual\_aqi\_by\_county\_2017.csv
  + Region Borders - <https://eric.clst.org/tech/usgeojson/>
    - gz\_2010\_us\_050\_00\_20m.json
  + Robert Wood Johnson Foundation - <https://www.countyhealthrankings.org/app/>
    - 2013 County Health Ranking Maryland Data - v1\_0.xls
    - 2014 County Health Rankings Maryland Data - v6.xls
    - 2015 County Health Rankings Maryland Data - v3.xls
    - 2016 County Health Rankings Maryland Data - v3.xls
    - 2017 County Health Rankings Maryland Data - v2.xls
    - 2013 County Health Ranking Virginia Data - v1\_0.xls
    - 2014 County Health Rankings Virginia Data - v6.xls
    - 2015 County Health Rankings Virginia Data - v3.xls
    - 2016 County Health Rankings Virginia Data - v3.xls
    - 2017 County Health Rankings Virginia Data - v2.xls
  + USCS - <https://www.cdc.gov/cancer/uscs/dataviz/download_data.htm>
    - BYAREA\_2015 (manually renamed)
    - BYAREA\_2016 (manually renamed)
    - BYAREA\_COUNTY\_2015 (manually renamed)
    - BYAREA\_COUNTY\_2016 (manually renamed)
* Pre\_Processing
  + preProcessing\_createCancerMasterDataFile
    - preProcessing\_createCancerMasterDataFile\_v\*.R
  + preProcessing\_createNonCancerMasterDataFile
    - preProcessing\_createNonCancerMasterDataFile\_v\*.R
    - countyHealthRankings
      * countyHealthRankings\_v\*.R

The following sections go through how these data folders are used to pre-process different segments of the visualizer.

*Maps*

The visualizer displays maps of the GWCC Catchment Area and its regions in several locations. This requires having latitude and longitude coordinates for each region. These are available for counties and states at this url: <https://eric.clst.org/tech/usgeojson/>.

The US Counties - 20m GeoJSON file was downloaded to the *Region Borders* folder. A copy of this was manually edited to keep just GWCC Catchment Area regions. This edited copy is found in the Shiny app as dmvGeojson.json.

Coordinates for DC wards were downloaded from the below url and added to dmvGeojson.json: <https://data.codefordc.org/dataset/dc-wards-map-overlay/resource/0004078a-8b07-463e-82d5-32921ef7c12f>.

*Pre-Processing - Cancer Data*

The *preProcessing\_createCancerMasterDataFile* folder contains a pre-processing file that takes the raw data files from the *USCS* and *DC Cancer Registry* folders to create the masterDataFile\_cancer\_countyWard.csv master cancer data file.

*Pre-Processing - Risk & Protective Factors*

The *preProcessing\_createNonCancerMasterDataFile* contains a pre-processing file that takes the raw data files from the *American Community Survey*, *EPA*, and *Robert Wood Johnson Foundation* folders to create the masterDataFile\_nonCancer\_countyWard.csv master risk & protective factors data file. In addition, it leverages the Census Bureau API to get additional information.

Shiny Architecture

The *Shiny\_Application* folder contains all the code for the visualizer. It is separated among DEV, TEST, and PROD environments.

*DEV Environment*

Each of the folders titled *Cancer\_Data\_Visualizer\_vXX.Y* constitutes a standalone Shiny application. The “XX” represents a major version and “Y” a minor version. Whenever you need to make a change to any of the files in the Shiny app, copy the latest Shiny app, increase the minor version by 1 in the folder name, and make your change in that folder. At the moment, I have just been increasing the major version by 1, so there is currently no meaning attached to the major version number. In the future, you may want to only increase the major version for a cohort of changes and allow a variable number of apps within a major version folder.

Per [RStudio’s recommendation](https://shiny.rstudio.com/articles/two-file.html), these apps have the following typical structure:

* ui.R - defines all front-end components
* server.R - defines all back-end components
* global.R - defines anything that can be referenced by the above 2 files
* www - contains any supporting files

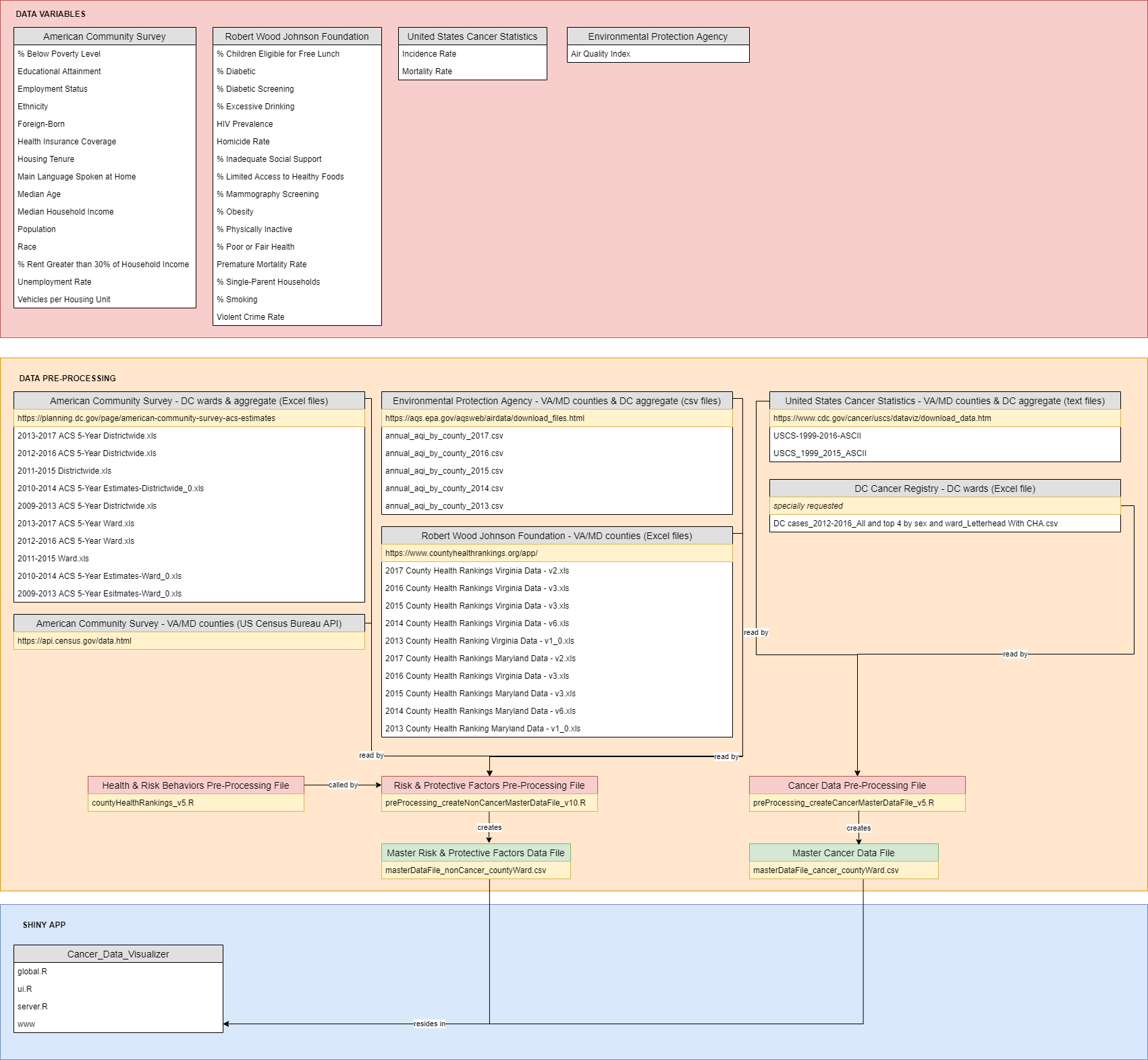
It is critical that these files and folder have these names.

After making a new version of the app, log that change in *www/developer/releaseNotes*.

Each of these apps in the DEV environment can be run locally by following these steps:

* ensure you have the folder *Cancer\_Data\_Visualizer\_vXX.Y* and all its contents
* open any of the global, ui, or server R files in RStudio
  + typically, RStudio’s *Source* pane has a “Run” button in the top right
  + after opening one of these files, RStudio will recognize the file as part of a Shiny app and this button will instead say “Run App”
* click “Run App”

The *DEBUG\_MODE* variable in the global R file can be used to aid development. When this value is set to TRUE, it allows you to type R code in the sidebar of the front-end and see the result in the RStudio console. In addition, it will expose some of the supporting files such as release notes, backlog, etc.

Below is a visual of how all data sources, pre-processing files, and visualizer files fit together:

Use [draw.io](https://www.draw.io/) to maintain this diagram.

*TEST & PROD Environments*

The higher environments are hosted on [shinyapps.io](https://www.shinyapps.io/). For this project, we have an account called “gwcancerdatavisualizer.” Since we have the Professional shinyapp.io account, any number of applications can be deployed here. Currently, we only have 2 applications.

First, <https://gwcancerdatavisualizer.shinyapps.io/cancer_data_visualizer/> contains the production app. This is being used by our end-users.

Second, <https://gwcancerdatavisualizer.shinyapps.io/cancer_data_visualizer_test/> contains a version of this app that is meant to act as pre-production. We can use this app for any purpose, whether it be ensuring deployments are running fine on shinyapp.io or to show someone a prototype.

Before you can deploy to shinyapps.io, you must configure the *rsconnect* R package to deploy to the *gwcancerdatavisualizer* account. Follow the steps in section 2.3.1 of the publicly available [shinyapps.io user guide](https://docs.rstudio.com/shinyapps.io/index.html) to do this.

Locally, you will notice in the TEST and PROD folders that in each folder there is one Shiny app that is just like the DEV apps except for one difference: instead of ui and server files, there is one app.R file that:

1. calls the global.R file
2. defines the ui object
3. defines the server object
4. concludes with *shinyApp(ui, server)*

Whenever it is time to deploy the latest DEV app, a specific sequence of steps must be performed. Below is an example of deploying from DEV to PROD:

1. replace the PROD global.R file with the DEV one
   1. open the PROD file and set DEBUG\_MODE to FALSE
2. open the PROD app.R file
   1. update the definition of *ui* with the contents of the DEV ui.R file
   2. update the definition of *server* with the contents of the DEV server.R file
3. delete the entire *www* directory in the PROD environment
   1. copy the *www* directory from the DEV environment over to the PROD environment
   2. this method ensures this directory is the exact same between environments
4. run *library(rsconnect)*
5. run *deployApp(appDir = "C:\\Users\\derek.funk\\Documents\\MSDS\\Capstone\\Cancer\_Data\_Visualizer\\Shiny\_Application\\3\_PROD\\Cancer\_Data\_Visualizer", account = "gwcancerdatavisualizer")*
   1. replace the path above with your own local path
   2. type “Y” in the RStudio console when prompted to update the already existing app

A similar sequence of steps would be taken when deploying from DEV to TEST.

Additionally, you could also deploy from TEST to PROD. In this case, it is safest to wipe out the contents of the 3\_PROD\Cancer\_Data\_Visualizer directory and then copy the contents from 2\_TEST\Cancer\_Data\_Visualizer\_TEST. Be careful not to change the names of these directories, as they are what appears in the app url.

In the case of a deployment error, consult chapter 9 of the [shinyapps.io user guide](https://docs.rstudio.com/shinyapps.io/index.html).

*Code Sharing*

To share Shiny app code from DEV with another individual, the entire contents of the Shiny app must be shared. Users with access to the *gwcancerdatavisualizer* account on shinyapps.io have the option to download the TEST or PROD code with the “Download Bundle” feature.