DataViz3

D21124026

2022-12-09

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

# -------------------------------------------------- #  
# -------------------------------------------------- #  
  
# Data Visualisation Assignment 2  
# Visualisations in R  
#  
# Student No. d21124026  
# Name : Ciaran Finnegan  
# TU060 Data Science MSc  
#  
# December 2022  
  
  
# Visualisation Three - Comparison of population   
# health against Socio-economic groups in Census 2011   
# and Census 2016 data for Irish counties  
# -------------------------------------------------- #  
# -------------------------------------------------- #  
  
# Additional packages to improve information display  
# on scatter plot graphs  
#install.packages("gghighlight")  
#install.packages("ggthemes")   
  
# load required libraries  
library(curl)

## Warning: package 'curl' was built under R version 4.2.2

## Using libcurl 7.64.1 with Schannel

library(readr)

## Warning: package 'readr' was built under R version 4.2.2

##   
## Attaching package: 'readr'

## The following object is masked from 'package:curl':  
##   
## parse\_date

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.2

library(sqldf)

## Warning: package 'sqldf' was built under R version 4.2.2

## Loading required package: gsubfn

## Warning: package 'gsubfn' was built under R version 4.2.2

## Loading required package: proto

## Warning: package 'proto' was built under R version 4.2.2

## Loading required package: RSQLite

## Warning: package 'RSQLite' was built under R version 4.2.2

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.2

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ purrr 0.3.5 ✔ forcats 0.5.2

## Warning: package 'tibble' was built under R version 4.2.2

## Warning: package 'tidyr' was built under R version 4.2.2

## Warning: package 'purrr' was built under R version 4.2.2

## Warning: package 'dplyr' was built under R version 4.2.2

## Warning: package 'stringr' was built under R version 4.2.2

## Warning: package 'forcats' was built under R version 4.2.2

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ readr::parse\_date() masks curl::parse\_date()

library(viridis)

## Warning: package 'viridis' was built under R version 4.2.2

## Loading required package: viridisLite

## Warning: package 'viridisLite' was built under R version 4.2.2

library(dplyr)  
library(scales)

## Warning: package 'scales' was built under R version 4.2.2

##   
## Attaching package: 'scales'  
##   
## The following object is masked from 'package:viridis':  
##   
## viridis\_pal  
##   
## The following object is masked from 'package:purrr':  
##   
## discard  
##   
## The following object is masked from 'package:readr':  
##   
## col\_factor

library(gghighlight)

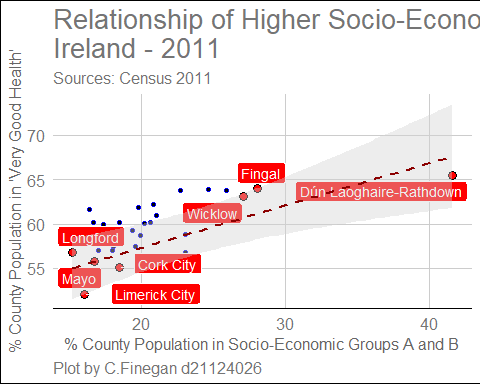
## Warning: package 'gghighlight' was built under R version 4.2.2

library(ggthemes)

## Warning: package 'ggthemes' was built under R version 4.2.2

library(stringr)  
  
  
  
# ---------------------------------------------------- #  
# Download Census Theme Data for 2016 for Irish counties  
# ---------------------------------------------------- #  
  
# Prepare URL string with location on GitHub of ZIP file with Census 2016 'Theme' data  
sGitHub\_Datasource1\_2016 <-"https://github.com/JackDaedalus/DataVizLabs/raw/"  
sGitHub\_Datasource2\_2016 <- paste(sGitHub\_Datasource1\_2016,"main/CA2/", sep = "", collapse=NULL)  
sGitHub\_Datafile\_2016 <- "SAPS2016\_CTY31.zip"  
sGitHub\_Datasource\_2016 <- paste(sGitHub\_Datasource2\_2016,sGitHub\_Datafile\_2016, sep = "", collapse=NULL)  
f2016CTY\_data <- sGitHub\_Datasource\_2016  
  
  
# Download zip file from from GitHub and extract 2016 Theme data for Irish counties  
temp\_3 <- tempfile()  
temp\_4 <- tempfile()  
source <- f2016CTY\_data  
temp\_3 <- curl\_download(url = source, destfile = temp\_3, quiet = FALSE)  
unzip(temp\_3, exdir = temp\_4)  
  
# Prepare location string of downloaded 2016 Census data CSV file   
f2016CensusData <- "\\SAPS2016\_CTY31.csv"  
f2016CensusData <- paste(temp\_4,f2016CensusData, sep = "", collapse=NULL)  
  
  
  
# ---------------------------------------------------- #  
# Prepare URL for download of Census Theme Data for 2011   
# for Irish counties  
# ---------------------------------------------------- #  
  
# Prepare URL string for Census Theme Data for 2011 for Irish counties  
f2011CensusData <- "https://www.cso.ie/en/media/csoie/census/documents/saps2011files/AllThemesTablesCTY.csv"  
  
  
  
  
  
# -----------------------------------------------------#  
# Prepare sequence of Census data to be downloaded and   
# read into dataframes for processing  
# -----------------------------------------------------#  
  
# Set up array of Census file names to be loaded in sequence  
arrCensusThemeFiles <- c(f2011CensusData, f2016CensusData)  
arrCensusThemeYears <- c('2011','2016') # 2011 data is downloaded first for manipulation  
  
  
  
# Set up dataframe array to hold the 2011 and 2016 Census Theme data  
# Both sets of data will undergo the same transformation before  
# being merged in advance of graph generation  
arrDFYrCountySocioThemes <- list() # start with empty array for data   
 # loaded from files  
arrDFYrCountySocioThemes\_Modified <- list() # array to store dataframes after   
 # the data wrangling process  
dfCensusGraph <- list()  
  
  
# -----------------------------------------------------#  
# Iterate through the 2011 and 2016 files and manipulate   
# the socio-economic data for visualisation  
# -----------------------------------------------------#  
#i <- 1  
for (i in 1:(length(arrCensusThemeFiles))) {  
  
   
 if (arrCensusThemeYears[i] == '2011') { # The 2011 Census Theme columns for   
 # Health data have a different format  
   
 # Select only the required Socio-economic and Health data   
 # Rename the columns to increase understanding of the data  
   
 # Read Census 2011 data from CSO website  
 arrDFYrCountySocioThemes[[i]] <- read\_delim(arrCensusThemeFiles[i],   
 show\_col\_types = FALSE) %>%   
   
 select(GEOGID, GEOGDESC, # Select the county identifier and the following...   
 T9\_2\_PA, # Socio-economic Group A  
 T9\_2\_PB, # Socio-economic Group B  
 T9\_2\_PT, # Total population in Socio-economic Groups  
 T12\_3VGT, # Population in 'Very Good Health'  
 T12\_3GT, # Population in 'Good Health'  
 T12\_3TT # Total Population in health census  
 ) %>%  
 rename(GroupA = T9\_2\_PA, # Rename Columns to improve readability  
 GroupB = T9\_2\_PB,  
 GroupsTotal = T9\_2\_PT,  
 VeryGoodHealth = T12\_3VGT,  
 GoodHealth = T12\_3GT,  
 HealthTotal = T12\_3TT)  
   
  
   
 } else {  
   
 # Read 2016 file - formats of Theme 12 column headings are different than those in 2011  
 arrDFYrCountySocioThemes[[i]] <- read\_delim(arrCensusThemeFiles[i],   
 show\_col\_types = FALSE) %>%   
 # Read Census 2011 data from CSO website  
 select(GEOGID, GEOGDESC, # Select the county identifier and the following  
 T9\_2\_PA, # Socio-economic Group A  
 T9\_2\_PB, # Socio-economic Group B  
 T9\_2\_PT, # Total population in Socio-economic Groups  
 T12\_3\_VGT, # Population in 'Very Good Health'  
 T12\_3\_GT, # Population in 'Good Health'  
 T12\_3\_TT # Total Population in health census  
 ) %>%  
 rename(GroupA = T9\_2\_PA, # Rename Columns to improve readability  
 GroupB = T9\_2\_PB,  
 GroupsTotal = T9\_2\_PT,  
 VeryGoodHealth = T12\_3\_VGT,  
 GoodHealth = T12\_3\_GT,  
 HealthTotal = T12\_3\_TT)  
   
  
 # Correct for Irish language fada in Dun-Laoighaire throwing a text error  
 arrDFYrCountySocioThemes[[i]][8, "GEOGDESC"] <- "Dun Laoghaire-Rathdown"  
   
 }  
  
  
  
 # ------------------------------------------------------------------#  
 # Data manipulation routines to add calculated field for   
 # each county for percentage of population in Social   
 # Class A ad B, and in 'Very God Health'  
 # ------------------------------------------------------------------#  
  
 # Create dataframe for year in array - 2011 or 2016   
 dfThisYrCountySocioThemes <- arrDFYrCountySocioThemes[[i]]  
   
 # Add Year Value as label to dataframe  
 dfThisYrCountySocioThemes$Year <- arrCensusThemeYears[i]  
   
   
 # Calculate Ratio by County of Population in Socio-economic groups A and B.  
 dfThisYrCountySocioThemes$GroupAB\_Pct <- ((dfThisYrCountySocioThemes$GroupA   
 + dfThisYrCountySocioThemes$GroupB)   
 / dfThisYrCountySocioThemes$GroupsTotal) \* 100  
   
   
 # Calculate Ratio by County of Population in 'Very Good' health.  
 dfThisYrCountySocioThemes$VGHealth\_Pct <- (dfThisYrCountySocioThemes$VeryGoodHealth  
 / dfThisYrCountySocioThemes$HealthTotal) \* 100  
  
  
   
 # Set up dataframe in array after data manipulation complete  
 arrDFYrCountySocioThemes\_Modified[[i]] <- dfThisYrCountySocioThemes  
   
   
   
 # -------------------------------------------------------------------#  
 # Prepare elements to generate and improve Scatter Plot  
 # rendering of relationship between Social Class A/B and  
 # 'Very Good Health'  
 # -------------------------------------------------------------------#  
   
 # Copy data to a dataframe with a shorter name to simplify  
 # the code generating the graphs  
 dfCensusGraph[[i]] <- arrDFYrCountySocioThemes\_Modified[[i]]  
   
 # Format graph titles - avoids distortion when rendering the graph  
 sGraphTitle <- "Relationship of Higher Socio-Economic Groups to Health in Ireland - "  
 sGraphSubTitle <- "Sources: Census "  
 # Add 'Year' for given dataframe in this loop  
 sGraphTitle <- paste(sGraphTitle, arrCensusThemeYears[i],   
 sep = "", collapse=NULL)  
 sGraphSubTitle <- paste(sGraphSubTitle, arrCensusThemeYears[i],   
 sep = "", collapse=NULL)  
   
   
 # These parameters work best for highlighting specific data points   
 # for county regions in the 2011/2016 Scatterplot graphs  
 UpperGroupAB\_Pct <- 27  
 LowerGroupAB\_Pct <- 16   
 LowerVGHealth\_Pct <- 56  
  
   
 # Generate Scatter Plot Grah  
 gg1 <- ggplot(data=dfCensusGraph[[i]], aes(x = GroupAB\_Pct, y=VGHealth\_Pct)) +  
 geom\_point(colour="black", size=3, shape=21, fill="red") +  
 # Highlight on Scatter plot the county areas at the upper and lower end of   
 # the socio-economic groups and health rating.  
 gghighlight(((GroupAB\_Pct>=UpperGroupAB\_Pct|GroupAB\_Pct<=LowerGroupAB\_Pct)   
 | VGHealth\_Pct<=LowerVGHealth\_Pct),   
 label\_key = GEOGDESC,  
 unhighlighted\_params   
 = list(colour = "blue",   
 fill="black",  
 size=1.5),  
 label\_params = list(fill="red",   
 colour="white",  
 size=3.75))+  
 # Label the graph  
 labs(x = "% County Population in Socio-Economic Groups A and B",   
 y = "% County Population in 'Very Good Health'",   
 # Format title to avoid distortion of the graph  
 title = str\_wrap(sGraphTitle, 60),  
 subtitle = sGraphSubTitle,   
 caption = "Plot by C.Finegan d21124026") +   
 # Add a dashed regression line and a confidence interval for relationship   
 geom\_smooth(method='lm', se=TRUE, linetype="dashed",  
 color="darkred", fullrange=TRUE, fill="light grey") +  
 # Use GDocs theme from ggplot themes library  
 theme\_gdocs()  
   
   
 print(gg1)  
   
  
}

## `geom\_smooth()` using formula = 'y ~ x'



## `geom\_smooth()` using formula = 'y ~ x'

