DATA 622: Machine Learning and Big Data: HW4 (Final Project)

Gabriel Campos

Last edited May 17, 2024

Assigment Description

Exploratory analysis and essay

Assignment

1. Choose a dataset You get to decide which dataset you want to work on. The data set must be different from the ones used in previous homeworks You can work on a problem from your job, or something you are interested in. You may also obtain a dataset from sites such as Kaggle, Data.Gov, Census Bureau, USGS or other open data portals.
2. Select one of the methodologies studied in weeks 1-10, and another methodology from weeks 11-15 to apply in the new dataset selected.
3. To complete this task:.
    a. Describe the problem you are trying to solve.
    b. Describe your datasets and what you did to prepare the data for analysis.
    c. Methodologies you used for analyzing the data
    d. What's the purpose of the analysis performed
    e. Make your conclusions from your analysis. Please be sure to address the business impact (it could be of any domain) of your solution.

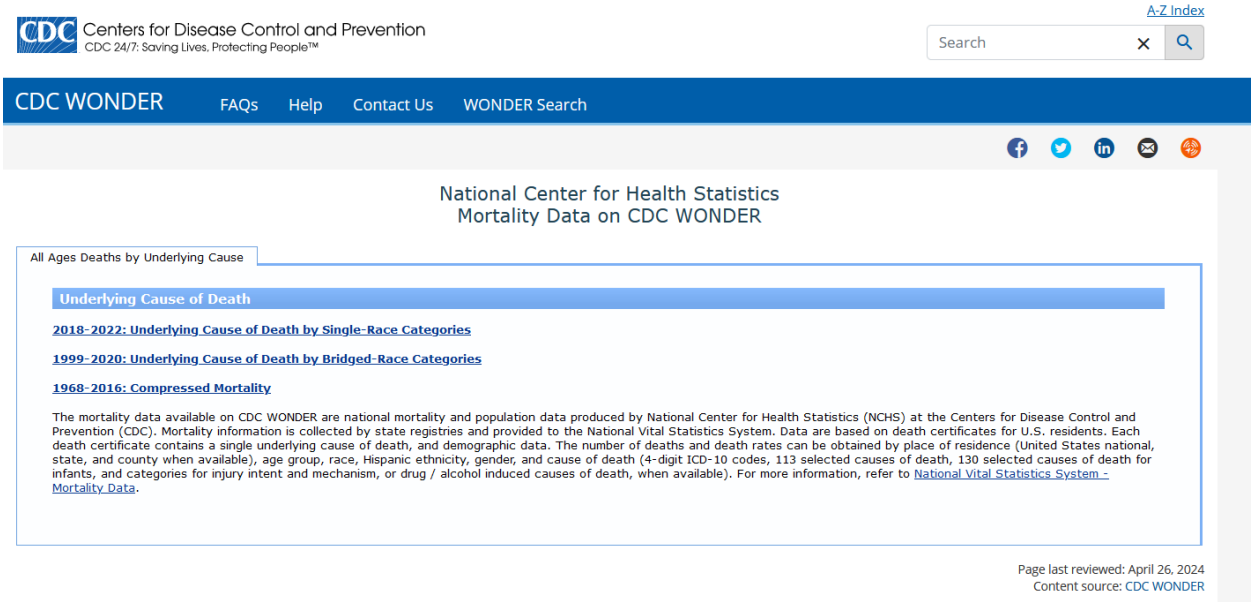
Deliverable

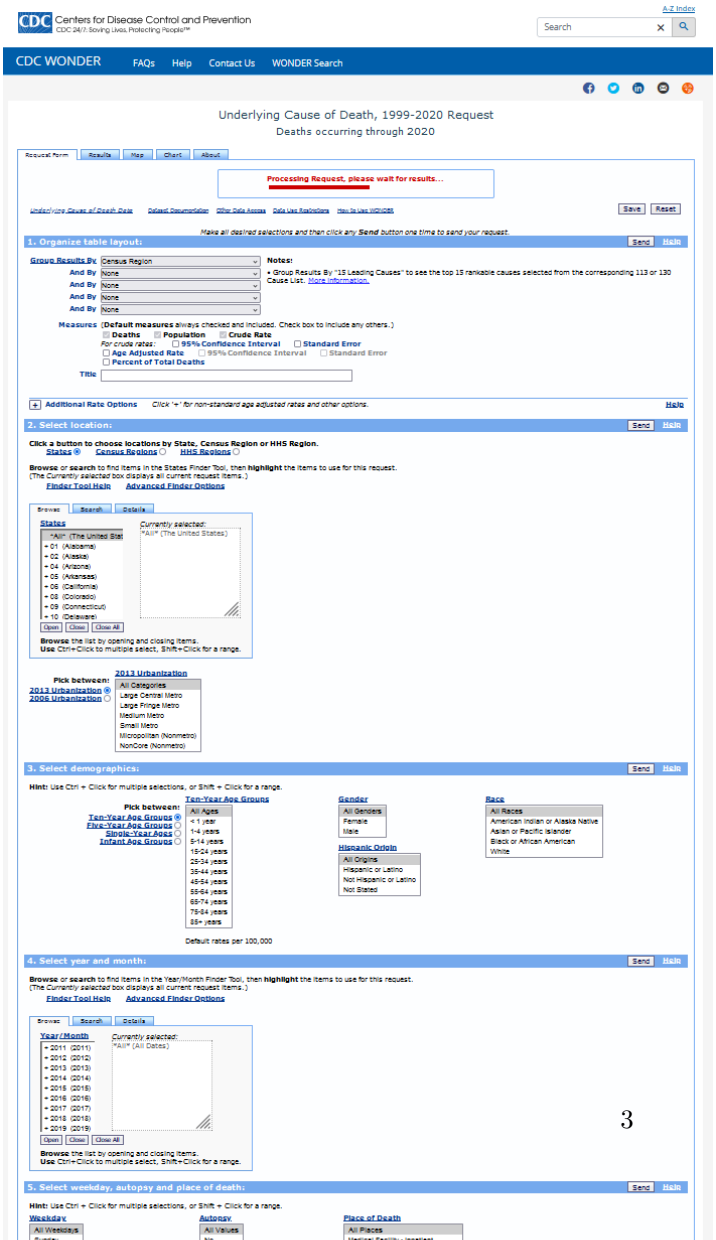
1. Your final presentation (essay or video) should include:
    1. The traditional R file or Python file and essay,
    2. An Essay (minimum 500 word document) or Video (5 to 8 minutes recording) Include the execution and explanation of your code. The video can be recorded on any platform of your choice (Youtube, Free Cam).

Libraries

```
library(Amelia)
library(car)
library(caret)
library(corrplot)
library(Cubist)
library(DataExplorer)
library(dplyr)
library(e1071)
library(earth)
library(forcats)
library(forecast)
library(fpp3)
library(gbm)
library(ggplot2)
library(kableExtra)
library(MASS)
library(mice)
library(mlbench)
library(party)
library(randomForest)
library(RANN)
library(RColorBrewer)
library(readxl)
library(rpart)
library(rpart.plot)
library(summarytools)
library(tidyr)
library(VIM)
library(earth)
library(randomForest)
```

Overview


A-Z Index
Search
Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™
CDC WONDER
FAQs
Help
Contact Us
WONDER Search
National Center for Health Statistics
Mortality Data on CDC WONDER
All Ages Deaths by Underlying Cause
Underlying Cause of Death
2018-2022: Underlying Cause of Death by Single-Race Categories
1999-2020: Underlying Cause of Death by Bridged-Race Categories
1968-2016: Compressed Mortality
The mortality data available on CDC WONDER are national mortality and population data produced by National Center for Health Statistics (NCHS) at the Centers for Disease Control and Prevention (CDC). Mortality information is collected by state registries and provided to the National Vital Statistics System. Data are based on death certificates for U.S. residents. Each death certificate contains a single underlying cause of death, and demographic data. The number of deaths and death rates can be obtained by place of residence (United States national, state, and county when available), age group, race, Hispanic ethnicity, gender, and cause of death (4-digit ICD-10 codes, 113 selected causes of death, 130 selected causes of death for infants, and categories for injury intent and mechanism, or drug / alcohol induced causes of death, when available). For more information, refer to National Vital Statistics System - Mortality Data.
Page last reviewed: April 26, 2024
Content source: CDC WONDER


Underlying Cause of Death, 1999-2020 Request
Deaths occurring through 2020
Processing Request, please wait for results...
1. Organize table layout:
2. Select location:
3. Select demographics:
4. Select year and month:
5. Select weekday, autopsy and place of death:

Load Data