Daniel Trevino dtrevi01@syr.edu
27 Feb 2019
IST 652 Scripting for Data Analysis

2017 CDC Mortality Data (Suicides)

Data Source: https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm

1. U.S. Data (.zip files)*
2. Unzipped data results in 1.2GB .DUSMCPUB file
3. Github user tommaho @ https://github.com/tommaho/VS13MORT.DUSMCPUB-Parser
    a. Parser for 2013 data was modified to extract all the text within the PUB file.

Data Exploration and Cleaning

1. Upon modification, attempted to put data into excel, which lead to failure due to row limitations in excel.
2. Opted to put the file in a .csv text file as seen below in the picture.

3. Read the file as a .csv into a pandas dataframe with the resulting columns

```
fileOutObj.write('Resident_Status, Education, Month_Of_Death, Sex, Age_Key, Age_Value, Age_Sub_Flag, Age_Recode_52, Age_Recode_27,
                 'Age_Recode_12, Infant_Age_Recode_22, Place_Of_Death, Marital_Status, DOW_of_Death, Data_Year, Injured_At_Work, '
                 'Manner_Of_Death, Method_Of_Disposition, Autopsy, Activity_Code, Place_Of_Causal_Injury,  ICD10, Cause_Recode_358
                 'Cause_Recode_113, Infant_Cause_Recode_130, Cause_Recode_39, Entity_Axis_Conditions, EAC1, EAC2, EAC3, EAC4, EAC5
                 'EAC6, EAC7, EAC8, EAC9, EAC10, EAC11, EAC12, EAC13, EAC14, EAC15, EAC16, EAC17, EAC18, EAC19, EAC20, ' + \
                 'Record_Axis_Conditions, RA1, RA2, RA3, RA4, RA5, RA6, RA7, RA8, RA9, RA10, RA11, RA12, RA13, RA14, ' + \
                 'RA15, RA16, RA17, RA18, RA19, RA20, Race, Race_Bridged, Race_Imputation, Race_Recode_3, Race_Recode_5, ' + \
                 'Hispanic_Origin, Hispanic_Origin_Recode\n')
```

4. After analysing all data needed for input, developed a data dictionary for the following:
    a. MannerOfDeath = 2:Suicide
    b. DeathDow = 1:Sunday 2:Monday 3:Tuesday 4:Wednesday 5:Thursday 6:Friday 7:Saturday
    c. DeathMonth: 10:October 11:November 12:December 01:January 02:February 03:March 04:April 05:May 06:June 07:July 08:August 09:September
5. The following columns was needed to obtain the data needed for further analysis of the data: ['Age_Value', 'sex', 'Place_Of_Death', 'Manner_Of_Death', 'Month_Of_Death', 'DOW_of_Death', 'Data_Year']
6. Created a new pandas dataframe with these columns extracted.
7. Only suicide deaths are required for the analysis, so a new data file called “Suicide.csv” was created:
    a. suicide = newData[newData[' Manner_Of_Death'] == 2] and output as a .csv

8. The Suicide.csv file contains the values for all recorded US suicides and will be the file used for the analysis.

Daniel Trevino dtrevi01@syr.edu
27 Feb 2019
IST 652 Scripting for Data Analysis

Data Analysis and Questions

1. The data contained within in Suicide.csv file contains the following information.

	Age	Sex	PlaceOfDeath	MannerOfDeath	DeathMonth	DeathDoW	Year
0	58	F	4	2	1	5	2017
1	46	F	7	2	1	1	2017
2	37	M	7	2	1	2	2017
3	51	M	4	2	1	3	2017
4	27	M	4	2	1	3	2017

2. All columns in the dataframe are integers except the ‘Sex’ Column. For the purpose of analysing the data, the Age, Sex, DeathMonth, and DeathDoW columns are used.
3. The following are the analysis questions output within MiniProject1V3.py file.
    a. What are the age differences and similarities between males and females.
    b. What is the mean age for males and females suicide for 2017?
    c. How many suicides by day of the week?
    d. What are the suicides by month count?
    e. What month to day combination results in the highest suicides?
4. The output of the questions, to include the data frame for analysis and a graph are within the .py file.

Program & Output File Description

1. The MiniProject1V3.py program outputs in the console both data analysis answers to the above questions as well as data frames. The data frames and plot of the data are also output to the local working directory which you put in. The output files are listed as analysis1, analysis2, analysis3, and analysis4. The analysis output corresponds to the questions listed above.
2. Each analysis file is a .csv with the header and values for analysis. These files are all grouped from the Suicide.csv. Additionally, the .csv files are plotted to illustrate the information contained within the analysis data frames and helps with interpreting the Data questions listed above.

Comparison Questions

1. Question a. above used the Age and Sex columns. The analysis1 output file just illustrates the findings. The findings as outlined in the .py print statement show males have several peak ages for suicide, 19, 28, and around 46 while women have a peak around 40 - 42.
    a. The average male suicide age for 2017 is 54 and the max age is 103.
    b. The minimum female age is 10, the average for 2017 is 51, and the max age is 102.
2. For analysis 2 the Sex and DeathDoW columns were used. The output of the comparison is both a .csv and a plot showing values.

Daniel Trevino dtrevi01@syr.edu
27 Feb 2019
IST 652 Scripting for Data Analysis

   a. The day in 2017 with the most Suicides is Monday with 1091 female and 3897 male deaths.
3. What month has the most suicides for males and females?
   a. The month of July 2017 contained the most male suicides and June 2017 contained the most female suicides.
4. What month to day combination results in the highest suicides?
   a. This is an interesting finding and best viewed in the analysis4_plot.png
   b. Males have the highest suicides on Mondays in May and on Mondays in July.
   c. Females have the highest suicides on Thursdays in June and on Mondays in July.
5. Further exploration is still required to analyze the combination of Place of death and the month to day combination from analysis4.