Milestone Report

Ish Gupta

December 21, 2016

An introduction to the problem (based on your earlier Capstone submissions).

National Heart, Lung and Blood Institute in United States followed a huge number of population for a period of 11 Years, to keep a record of their lifestyle Habits, Socio economic status, Education, and Health. Since, with the Survey- people were in different age groups, not everyone could survive the period of 11 years. For any individual who could not live longer, was recorded the cause of Death. From the population, there are a total of 113 causes of death, where there are different causes relating to Health- if they had any pre- existing diseases, Economical wellness- how much did they earn, to what Class of the SocioEconomic band in society did they belong to?; Educational Background.

The aim is to find out the correlations of Deaths, with their respective characteristics of Life, where the Cause of Death is SUICIDE.

A deeper dive into the data set:

The dataset provides a total of 1835072 surveys with attributes providing below details of a Person:

S. No.	Variable	Description
1	age	Age of the Person
2	race	Race
3	sex	Gender
4	ms	Marital status at time of survey.
5	adjinc	“Inflation Adjusted Income”- Family income corresponsing to the CPI.“
6	educ	Highest grade completed at the time of the interview.
7	reltrf	Relationship to reference person within the household.
8	ind	Job specific industrial classification codes.
9	esr	Recoded employment status.
10	inddea	“Death Indicator”- An indicator variable identifying decedents in the file. Deaths occurred within the associated follow-up period for this file after the survey interview.“
11	cause113	A recode of the underlying cause of death
12	follow	The length of follow-up period in days.
13	histatus	“Health Insurance Status”- Indicates whether person had health insurance coverage any time in the calendar year prior to the interview.“
14	hitype	“Health Insurance Type”- This variable summarizes, in broad categories, the type of health insurance coverage that the person had at any time in the previous calendar year.“
15	stater	State of Residence. First digit of STATER is the Census Bureau division code. The second digit is the state within the division code.

S. No.	Variable	Description
16	tenure	Variable defines the type of ownership of the residence. All members of the household receive the same values.
17	citizen	This variable indicates the person’s citizenship at the time of survey.
18	health	This variable is the response to the question: “Would you say that [person]’s health in general is”.
19	smok100	“Smoked More than 100 Cigarettes”- Has [person] smoked at least 100 cigarettes in his/her lifetime“
20	agesmk	“Age Started Smoking”- How old was [person] when he/she started smoking cigarettes fairly regularly?“
21	smokstat	“Cigarette Smoking Status”- indicates the general quantity of cigarette smoking.“
22	hisp	Hispanic origin classifies all persons by Mexican, Hispanic (not Mexican) or Non-Hispanic (Not Mexican or Hispanic) origin.
23	pob	Place of Birth
24	hhnum	The number of persons residing in the household at the time of the interview.
25	occ	“4 Digit Occupation Code”- Job specific 2000 occupational classification codes“
26	majocc	“Major Occupation Code”- Major occupation classification recode of 2000 specific occupational codes.“
27	majind	Major industry classification recode of 2007 specific industry.
28	urban	Person lives in an Urban area or Rural.
29	dayod	The day of the week on which the decedent died.
30	hosp	“Hospital Type”- place of death.“
31	hospd	“Hospital Death Indicator”- Indicates the location of death relative to a hospital. Response is determined from the death certificate. Certain states do not have this item on their certificate. For these states, a code of missing is assigned.“
32	vt	Indicates whether person was U.S. veteran.
33	povpct	“Income as Percent of Poverty Level”- This variable is defined by taking the family income, adjusted for inflation to 1990 dollars, and comparing it to the 1990 defined poverty level.“
34	rcow	“Recoded Class of Worker”- A recode of the detailed class of worker as derived from the type of work performed “
35	indalg	“Indicator for Algorithmic Death”- indicator classifies death status based only on the computerized screening algorithm from both recent and previous matches to the NDI. If (INDDEA = 1) and (INDALG=1) then INDMORT = 1. Else INDMORT = 0.“
36	smokhome	“Rules for Smoking Cigarettes in the Home”- Which statement best describes the rules about smoking [cigarettes] in your home?“

S. No.	Variable	Description
37	curruse	“Currently Use Smokeless Tobacco”- 5-digit character string in which each digit represents the current use response for other types of other tobacco use. Reading the digits from left to right, the individual digit components are:
38	everuse	“Ever Use Smokeless Tobacco”- have you ever used tobacco other than the smoking of cigarettes; a 5-digit character string in which each digit represents a have used response for the various types of other tobacco uses that were surveyed.
39	wt	Weights were obtained by raking age-sex-race group totals by state totals for each survey.
40	smsast	Is a household located in an SMSA or not?

What important fields and information does the data set have?

Cause of Death

Predictors which are correlated and could also be Causal for the Cause of Death involving Gender, Education, Health, Economical wellness, Location.

S.No.	Variable	Description
1.	age	Age
2.	race	Race
3.	ms	Marital Status
4.	educ	Education
5.	inddea	Dead or Alive
6.	adjinc	Family Income
7.	ind	Job specific industrial code
8.	histatus	Had insurance or not
9.	stater	State of residence
10.	health	Quality of Health
11.	agesm	Age Started Smoking

What are its limitations i.e. what are some questions that you cannot answer with this data set?

Human behavior is the most unpredictable thing. Even if, there is find to be correlation, and some level of Causality- we cannot be sure of the accuracy this model will be able to provide. It will be helpful in getting to know the factors which had an affect on the people part of this Survey, and a chance to predict by splitting the Data into Training and Test sets.

What kind of cleaning and wrangling did you need to do?

1. Remove the Id variable, since it is not helpful as a predictor.

2. removed ‘inndea’, since it indicates if a person is Alive or not. But, we are already studying Suicides, so person is actually, no more.
3. Impute the missing values with medians.
4.

Any preliminary exploration you’ve performed and your initial findings.

Plotted individual respective histograms, to discover most of the Suicides are committed by people: a) with ‘Excellent’ health. b) who are native and born in US, but 87% of the total population do not have their Citizenship recorded. c) who have Education till “Completed H4”. d) who were employed in the labor force, and least by people employed in the labor force, but were Absent from work. e) who are White. f) having Social Security Number on the CPS record. g) living in Rural areas. h) of Male population. i) born in US. j) who are Married. k) who are Ref person with other relatives in household. (RELTRF) l) where the number of people residing the Household is 2. m) who are living a House owned by them. n) who work in Manufacturing (Nondurable Goods) o) who work in Private industry, and minimum by the people working without pay, and those who have never worked. p) who had insurance, provided by their Employer. q) were a Smoker, and started smoking between an age of 12 to 18, with a median of 16. Interestingly, 36% have never smoked, and 39% are a regular smoker. r) who do not consume Tobacco of any kind.

Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?

Will try to find correlation between the variables which seem to have a majority of people committing Suicides.