**Statistics 6740**

**CLASS PROJECT**

**Due Date: November 23, 2015**

**Overview**

The mission of the Ohio Department of Veterans Services (ODVS) is to empower “the veteran and military community through service, support, and economic opportunity.” As one means of accomplishing its mission, the ODVS collaborates with a network of County Veteran Services Offices (CVSOs), with an office in each Ohio county. Each office is run by an appointed commission whose members are veterans from the major veteran service organizations. ODVS conducts outreach to veterans through the CVSOs, coordinates programs and operations with the U.S. Veterans Administration (VA), and works with the Ohio State Legislature on legislation affecting veterans and with other state agencies on programs that affect veterans.

One of the services offered by the county offices is assistance for veterans in completing disability and pension claims. Disabled veterans can receive monthly disability compensation, with the amount varying depending on a “percent disabled” rating they receive when their application is adjudicated by the regional VA office in Cleveland.

In a 2006 review, the Institute for Defense Analysis found that the percent of veterans who receive disability compensation in Ohio was the lowest among all states. Since that review, ODVS and the County Veterans Service Commission network have worked to increase that percentage, with some success. As part of their effort, they used information from the Department of Veterans Affairs to track veterans who were connected to Federal VA disability compensation by each county office. In the past, county performance in disability connection had been judged against similarly-sized counties without regard to other factors. However, based on a preliminary examination of the data from the various counties, ODVS believed that economic and demographic situations in each county were likely affecting the rate at which the county office is able to connect veterans.

ODVS undertook a project to develop a statistical model to predict the percentage of veterans in each county who were receiving disability compensation to determine if there were factors that predicted higher percentages that perhaps could be modified within counties whose percentages were low. In order to develop this model, ODVS obtained a large number of potential predictive factors from various sources, including the Department of Veterans Affairs, the American Community Survey, the Ohio Department of Job and Family Services, the Social Security Administration, and the CVSOs. The predictive factors are collected on the county level, resulting in 88 observations.

These data have been combined into an Excel spreadsheet entitled OH County Data, which has been placed on the class Carmen site in the Class Project folder. The file contains several tabs, each with a different type of data. Each tab contain the county name and a fips identifier for the county (GEO\_ID). A description of the remaining variables in each tab are as follows:

* General
  + Totpop – total county population
  + Area\_sqmi – county area in square miles
  + pop\_dens – population density (persons per square miles)
  + pct\_urb – percent of population in urban areas
  + rel\_adhere – percent religious adherence
  + pct\_usnative – percent U.S. natives
  + pct\_naturalized – percent naturalized citizens
  + pct\_fem – percent female
  + Urbanicity – urbanicity index for county (0 = very low, 1 = low, 2 = medium, 3 = high)
  + App\_ind – indicator of whether the county is in the Appalachian region
* Workforce
  + pct\_unemployed1 – percent unemployed
  + pct\_unemployed2 – percent unemployed
  + numcivlabor – number of civilians in the labor force
  + numdisablework – number of disabled workers
  + pct\_wf\_disable – percent of workforce that is disabled
  + pct\_wf\_ba – percent of the workforce with a BA degree
  + pct\_wf\_somecoll – percent of the workforce with some college (but not a degree)
  + pct\_wf\_higher – percent of the workforce with a post-graduate degree
* Health
  + health\_score – average health score for the county
  + pb\_health\_rk – public health rank among all counties
  + pct\_pop\_disable – percent of the population that is disabled
  + pct\_healthins – percent of the population with health insurance
* Housing
  + pct\_occupied – percent of housing units that are occupied
  + pct\_renters – percent of housing units containing renters
  + med\_home\_value – median home value
  + pct\_samehouse – percent of households that have been in the same house for the last year
* Family
  + pct\_fam – percent of households that are families
  + Pct\_fam\_child – percent of households that are families with children
  + pct\_married – percent of households that include a married couple
  + pct\_married\_child – percent of households with a married couple and children
  + pct\_fem\_head – percent of households with a female head
  + pctfemhead\_child – percent of households with a female head and children
  + med\_fam\_size – median family size
  + pct\_divorced – percent households with divorced individuals
* Race
  + pct\_multrace – percent of individuals who are multi-racial
  + pct\_white – percent of individuals who are White
  + pct\_afraf – percent of individuals who are African American
  + pct\_natvam – percent of individuals who are Native American
  + pct\_asian – percent of individuals who are Asian
  + pct\_pacisl – percent of individuals who are Pacific Islanders
  + pct\_hisp – percent of individuals who are Hispanic
* Income
  + med\_inc – median household income
  + pct\_inc\_lt10 – percent of households with income less than $10,000
  + pct\_inc\_10\_15 – percent of households with income between $10,000 - $15,000
  + pct\_inc\_15\_25 – percent of households with income between $15,000 - $25,000
  + pct\_inc\_25\_35 – percent of households with income between $25,000 - $35,000
  + pct\_inc\_35\_50 – percent of households with income between $35,000 - $50,000
  + pct\_inc\_50\_75 – percent of households with income between $50,000 - $75,000
  + pct\_inc\_75\_100 – percent of households with income between $75,000 - $100,000
  + pct\_inc\_100\_150 – percent of households with income between $100,000 - $150,000
  + pct\_inc\_150\_200 – percent of households with income between $150,000 - $200,000
  + pct\_inc\_200up – percent of households with income over $200,000
  + pct\_pov pct\_pov\_65 – percent of population over 65
* Age
  + pct\_0\_5 – percent of the population less than 5 years old
  + pct\_5\_9 – percent of the population aged 5 – 9 years
  + pct\_10\_14 – percent of the population aged 10 – 14 years
  + pct\_15\_19 – percent of the population aged 15 – 19 years
  + pct\_20\_24 – percent of the population aged 20 – 24 years
  + pct\_25\_34 – percent of the population aged 25 – 34 years
  + pct\_35\_44 – percent of the population aged 35 – 44 years
  + pct\_45\_54 – percent of the population aged 45 – 54 years
  + pct\_55\_59 – percent of the population aged 55 – 59 years
  + pct\_60\_64 – percent of the population aged 60 – 64 years
  + pct\_65\_74 – percent of the population aged 65 – 74 years
  + pct\_75\_84 – percent of the population aged 75 – 84 years
  + pct\_85up – percent of population aged 85 and older
  + pct\_18up – percent of population aged 18 and older
  + pct\_21up – percent of the population aged 21 and older
  + pct\_65up – percent of the population aged 65 and older
* Education
  + num\_school – number of individuals in school
  + pct\_ed\_lt9 – percent of individuals with less than a grade 9 education
  + pct\_ed\_9\_12 – percent of individuals with some high school education (no degree)
  + pct\_hsgrad – percent of individuals with a HS degree
  + pct\_ed\_somecoll – percent of individuals with some college education
  + pct\_ad\_assoc – percent of individuals with an associate degree
  + pct\_ed\_collgrad – percent of individual with a college degree
  + pct\_ed\_gradprof – percent of individuals with a graduate or professional degree
* Veterans
  + VetPop2011 – number of veterans
  + vet\_dens – veteran density (number of veterans per total population)
  + numvet65\_84 – number of veterans aged 65 to 84 years
  + pctvet65\_84 – percent of veterans aged 65 to 84 years
  + numvet85up – number of veterans aged 85 and older
  + pctvet85up – percent of veterans aged 85 and older
  + numvet17\_44 – number of veterans between the ages of 17 and 44
  + pctvet17\_44 – percent of veterans between the ages of 17 and 44
  + numvet65up – number of veterans aged 65 and older
  + pctvet65up – percent of veterans of aged 65 and older
  + numvet44\_64 – number of veterans between the ages of 44 and 64
  + pctvet44\_64 – percent of veterans between the ages of 44 and 64
  + pctvet17\_64 – percent of veterans between the ages of 17 and 64
  + numvetfem – number of female veterans
  + pctvetfem – percent of female veterans
  + med\_inc\_vet – median income for veterans

Your assignment is to perform an exploratory data analysis of this dataset and prepare a summary report using the guidance provided below.

**Assignment Details**

Perform the following tasks:

1. Download the Excel file from the Carmen site and read the data from each tab into SAS. Make sure that all variables have the appropriate type (character, numeric).
2. Create labels for all variables and formats for numeric variables that are categorical. The labels should be the variable descriptions provided above.
3. Merge all of the datasets. Make sure that you have one observation per county.
4. The State of Ohio is often divided into five regions. The map on the following webpage shows the counties that are in each of the five region. Add the region for each county to the database. <https://landsdssustainable.files.wordpress.com/2013/10/new-ohio-regions-map-350dpi-copy.jpg>.
5. Create the following new variables (with appropriate labels and formats):
   1. Affluence – if median household income is greater than $60,000 then “High”; if median household income is less than $40,000 then “Low”; otherwise, “Medium”;
   2. County size – if square miles is less than 425 then “Small”; if square miles is greater than 550, then “Large”; otherwise medium.
6. Select one continuous variable from each of the 10 Excel worksheets (your choice, with the exception of median household income and county size, which you should not use).
7. Perform the following exploratory data analyses
   1. Single variable analyses:
      1. Counts of counties by region, urbanicity, affluence, county size, Appalachian location;
      2. Summary statistics (select 5 or 6 useful statistics) for each of your selected continuous variables
   2. Two-variable analyses:
      1. Comparisons between regions: urbanicity level, affluence, Appalachian location;
      2. Comparisons between urbanicity levels: affluence;
      3. Comparisons of your selected continuous variables between region, urbanicity, affluence, and Appalachian location (separately);
      4. Comparisons of each pair of your selected continuous variables (use PROC GSCATTER);
   3. Multi-variable analyses:
      1. Summary statistics for your selected continuous variables by region and affluence;
      2. Comparison of your summary statistics by urbanicity and affluence;
      3. Select 3 of your continuous variables and compare each pair between regions;
8. Write and a report (in Microsoft Word) submit it to the Carmen Project Dropbox that includes:
   1. Title page
      1. Report title
      2. Student name, Course information
   2. Introduction
      1. Background
      2. Data sources
      3. Objectives of analysis
   3. Database description
      1. Contents
      2. Conversion of raw data to SAS dataset
   4. Summary of Exploratory Data Analysis
      1. Tables
      2. Graphics
      3. Text describing analysis and result
   5. Appendix – SAS code

**Notes**

* + - 1. Each of the analyses above should produce a table or figure (or both) to put in your report. You should make sure to have at least one each of the following:
         1. frequency table,
         2. summary statistics table,
         3. scatterplot,
         4. bar chart, and
         5. pie chart.
      2. You are free to add any other exploratory analyses that interest you.
      3. You are free to add any statistical analyses, although none are required (please restrict it to the set of variables you are using in the rest of the analysis/report).
      4. The Introduction and database description can borrow heavily from the material at the beginning of the assignment, although I expect you to do more than simply cut and paste – use some of your own language.