**CSC423: Data Analysis And Regression / CSC 324: Data Analysis & Statistical Software II**

**Assignment-1** | **Total points: 15 for CSC 324 and CSC 423**

**Due Date: 01/23/2018 by 11:59 pm**

**PROBLEM 1 [10 pts] – to be answered by everyone**

The filevoting\_1992.txt attached to this assignment provides data acquired from census records selected counties in the U.S. who voted in 1992 elections. The data show

County – Name of the county

Pct\_Voted – Percentage of people voted

MedianAge – Median age of the voters in that county

MeanSavings – Mean savings in U.S. Dollars in that county

Pct\_Poverty – Percentage of people living in poverty in that county

PopulationDensity – Population density (Population divided by square miles) in that county

Gender – Dominant gender of the people voted in that county

***Use SAS to compute the analysis below. All the functions are in either the code for the Lab Session-1 we did in class (see code that was posted on D2L). This is the first assignment, and for many of you it may be the first time you use SAS outside of the first lab session. So if you run into an error, post a message on the discussion board or contact me. Make sure to include your code in the message.***

In this exercise you are asked to get the data into a SAS dataset and perform basic exploratory analysis of the data to analyze the characteristics of people voted.

1. Open the dataset and examine the data. Answer the following:
   1. How many Observations are there?
      1. 884
   2. How many fields are there?
      1. 7
   3. Which fields are numerical?
      1. Pct\_Voted
      2. MEdianAge
      3. MeanSavings
      4. Pct\_Poverty
      5. PopulationDensity
   4. Which fields are text?
      1. Gender, County
2. Write the SAS code to create the SAS dataset using either IMPORT or INFILE statement. If you are using INFILE statement, pay attention to the text fields while writing your code.

Title 'Voting\_1992';

**PROC** **IMPORT** datafile='S:\LabSession1\voting\_1992.txt' out=voting replace;

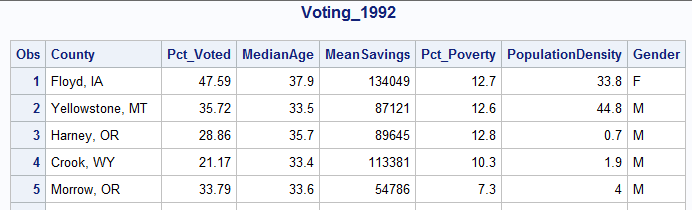
delimiter='09'x;

getnames=YES;

datarow=**2**;

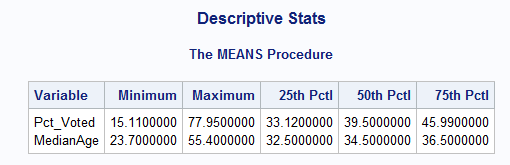
**RUN**;

1. Run a PROC PRINT to print your dataset in SAS. Do a print screen, to copy and paste the first 5 observations of the output.



1. What is the 5-point summary numbers for percentage of people voted and median age? The 5-point summary numbers are min, max, median or 50% percentile, Q1 and Q3.

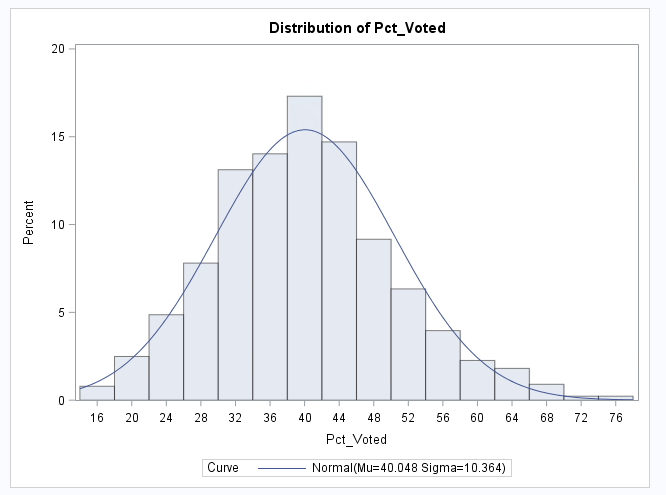
Include the output. Discuss your findings.



For Pct\_Vote, the range is 62.84.

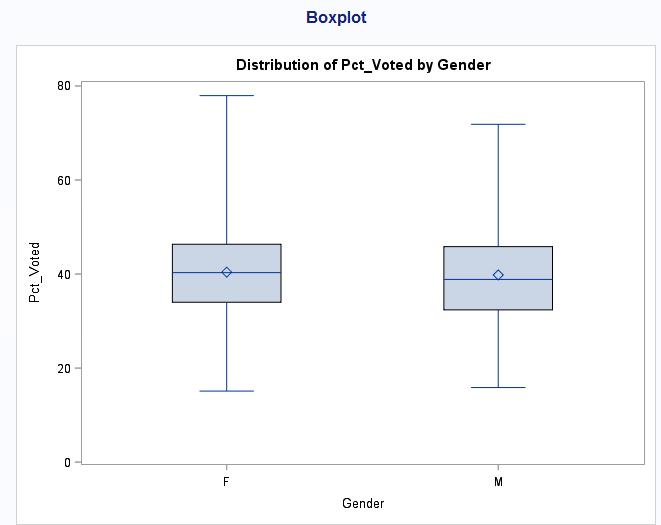
For MedianAge the range is 31.7.

1. Create a histogram to analyze the percent people voted. Include the histogram output. Using the histogram and the 5-point summary from the previous question, analyze the histogram. Discuss your findings. Also, is it normal, or skewed; do you see outliers?



The distribution is normal, and there are no outliers. The mean is 40.048 and the median is 39.50. Since the mean and median are close, this shows the distribution is normal.

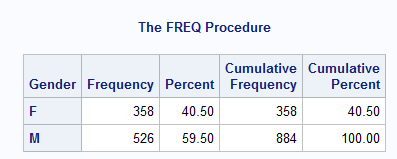
1. Create a boxplot to analyze percentage of people voted by gender. Include the output. What can you say about the gender and voting patterns? Discuss your findings using the boxplot.



The median and mean for F and M seem to be very similar, but the max for F is larger than M. It appears that the both have the same minimum percent voted.

1. What is the gender breakdown in this dataset? (Hint: use PROC FREQ). Include the output. Which is the predominant gender in this dataset?

Males are the predominant gender in the dataset.



1. Copy and paste your FULL SAS code into the word document along with your answers.

Title 'Voting\_1992';

**PROC** **IMPORT** datafile='S:\LabSession1\voting\_1992.txt' out=voting replace;

delimiter='09'x;

getnames=YES;

datarow=**2**;

**RUN**;

**PROC** **print**;

**run**;

Title 'Descriptive Stats';

**PROC** **means** min max p25 p50 p75 ;

var pct\_voted medianage;

**run**;

Title 'Histogram of Pct\_vtd';

**PROC** **univariate** normal;

var pct\_voted;

histogram / normal (mu=est sigma=est);

**run**;

Title 'Boxplot';

**Proc** **sort**;

by gender;

**run**;

**proc** **boxplot**;

plot pct\_voted\*gender;

**run**;

\*freq test;

Title 'Frequency';

**proc** **freq**;

tables gender;

**run**;

**PROBLEM 2 [5 pts] – to be answered by everyone**

Examine the 2 code segments and answer the following questions.

***Code-1***

**data** cpu;

infile "cpudat.txt" delimiter=',';

input time line step device;

**run**;

***Code-2***

**proc** **import** datafile="cpudat2.csv" out=cpu\_imp replace;

delimiter=' ';

DATAROW=**1**;

getnames=YES;

**run**;

*Note:*

*See link if you don’t know what a file extention is:* [*https://www.lifewire.com/what-is-a-file-extension-2625879*](https://www.lifewire.com/what-is-a-file-extension-2625879)

1. **The datafile name used in Code-1 is** cpudat.txt
2. **The datafile name used in Code-2 is** cputdat2.csv
3. **SAS dataset name for Code-1 is** cpu
4. **SAS dataset name for Code-2 is** cpu\_imp
5. **The delimiter used in Code-1 is** comma
6. **The delimiter used in Code-2 is** space
7. **The datafile extension of Code-1 is** .txt
8. **The datafile extension of Code-2 is** .csv
9. **Tue or False? “Code-1 uses import statement while Code-2 uses infile statement”** False
10. **Tue or False? “Code-1 and Code-2perform the same type of function – i.e. they both allow you to get the data written into SAS dataset”** True