**Kubra Iqbal**

**assignment 1**

**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: 04/03/2018 by 11:59 pm**

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

The file voting\_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? 884
   2. How many fields are there? Seven
   3. Which fields are numerical? Pct\_Voted, MedianAge, Pct\_Poverty and Population Density
   4. Which fields are text? County, Gender
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 "observations";

**PROC** **IMPORT** datafile="voting\_1992.txt" out=voting replace;

delimiter='09'x;

getnames=YES;

datarow=**2**;

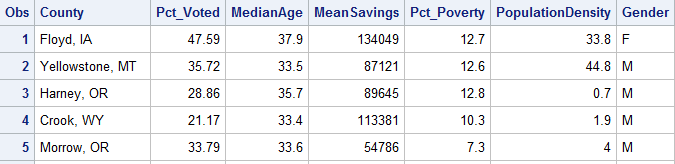
**RUN**;

\*print dataset gpa;

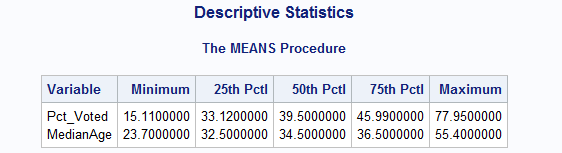
**PROC** **PRINT**;

**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.

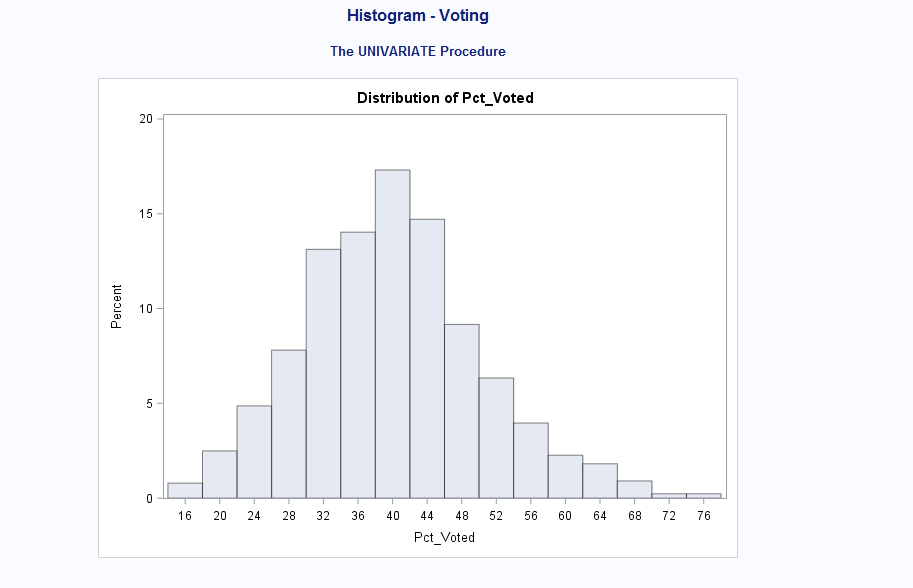


1. What is the 5-point summary numbers for percentage of people voted and median age? The 5-point summary numbers are min, max, median or 50% percentile, Q1 and Q3. Include the output. Discuss your findings.



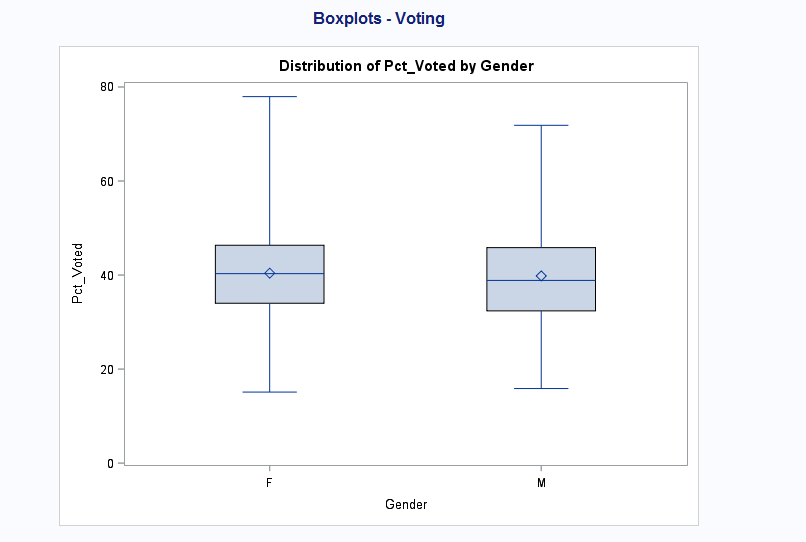
The above result shows that as the Median age increases, the percentage of people that have voted have gradually increased with it. It has increased from the Lower Quartile to Higher Quartile.

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 percentage of people who voted increases and then decreases as shown in the Histogram. The highest it goes is up to 40 and then comes down gradually. As the age increases – the percentage of people keeps increasing until approximately 18 when the percentage of people who voted starts to decrease gradually. The histogram is symmetric.

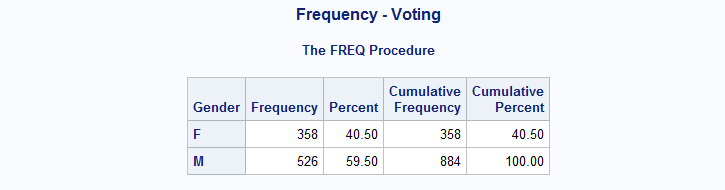
1. Create a boxplot to analyze percentage of people voted by gender. Include the output. What can you say about the gender and voting patterns? Discuss your findings using the boxplot.



For Female – the boxplot shows that the minimum is around 18 and Q1 is 38 approximately. The median is 40. Q3 is approximately 45 and maximum is around 78.  
For Male – The boxplot shows that minimum is around 18 approximately, Q1 is 35 and median is 38 and Q3 is 45. The max for this boxplot is 70.

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

Male – predominant in the gender set.



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

Title "observations";

**PROC** **IMPORT** datafile="voting\_1992.txt" out=voting replace;

delimiter='09'x;

getnames=YES;

datarow=**2**;

**RUN**;

\*print dataset gpa;

**PROC** **PRINT**;

**RUN**;

TITLE "Descriptive Statistics";

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

var Pct\_Voted MedianAge;

**RUN**;

TITLE "Histogram - Voting";

**PROC** **UNIVARIATE** ;

var Pct\_Voted;

histogram;

**RUN**;

TITLE "Boxplots - Voting";

**PROC** **SORT**;

BY Gender;

**RUN**;

**PROC** **BOXPLOT**;

PLOT Pct\_Voted\*Gender ;

**RUN**;

TITLE "Frequency - Voting";

**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 extension 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 cpudat2.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 “,”**
6. **The delimiter used in Code-2 is “ “**
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-2 perform the same type of function – i.e. they both allow you to get the data written into SAS dataset” --- true**