**PROBLEM 1 [10 pts] – To be answered by everyone**The file election.txt attached to this assignment provides data acquired from census records of selected counties in the U.S. who voted in the elections. The data shows

County PctVoted MedianAge MeanIncome PctUnemployment PopulationDensity Gender

– Name of the county

– Percentage of people voted

– Median age of the voters in that county

– Mean income in U.S. Dollars in that county

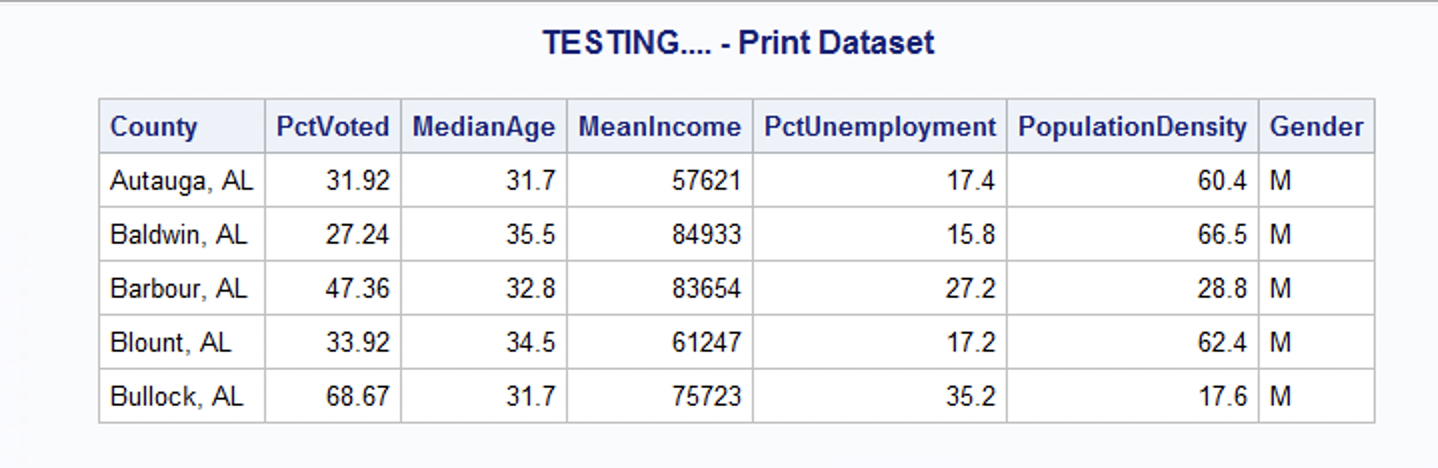
– Percentage of people unemployed in that county

– Population density (Population divided by square miles) in that county – Dominant gender of the people voted in that county

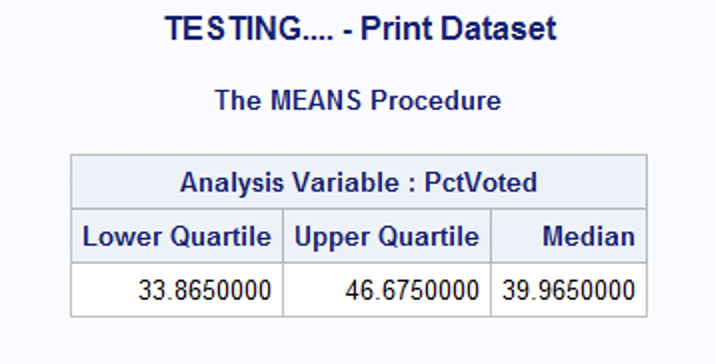
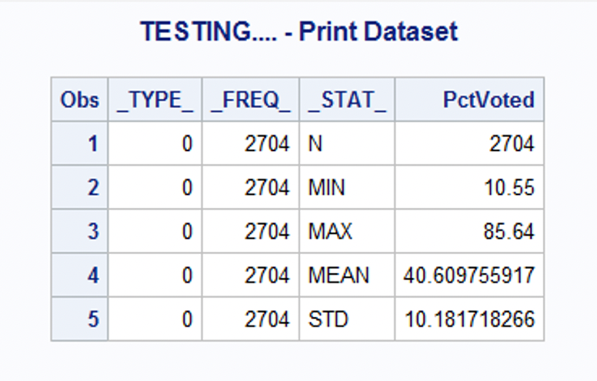
***Use SAS to compute the analysis below. All the functions are in either the code for the Lab Session-1 we did in class (see code that was posted on D2L). This is the first assignment, and for many of you it may be the first time you use SAS outside of the first lab session. So if you run into an error,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

* a)  Open the dataset and examine the data. Answer the following:
* 1. How many Observations are there? **2704 observations**  
  2. How many fields are there? **7 fields**  
  3. Which fields are numerical? **5 fields**4. Which fields are text? **County and Gender**
* b)  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.
* (Prints all observations)
* Graphical user interface, application

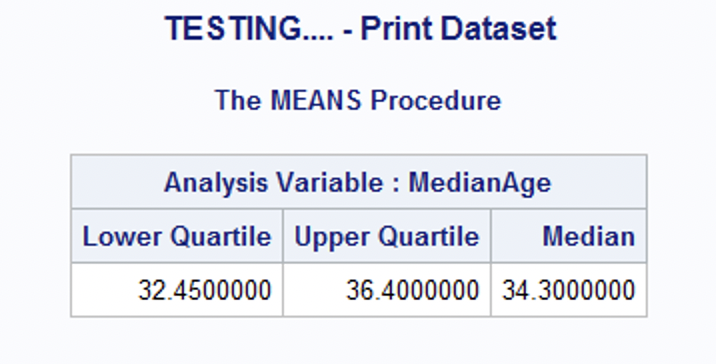
  Description automatically generated with medium confidence
* c)  Run a PROC PRINT to print your dataset in SAS. Do a print screen, copy and paste the first 5 observations of the output.
* (Prints first 5 observations)
* 
* d)  What are the 5-point summary numbers for percentage of people voted, median age and percent unemployment? The 5-point summary numbers are min, max, median or 50% percentile, Q1 and Q3. Include the output. Discuss your findings based on the values you see.

**Percentage of People Voted 5-point summary:**



PctVoted: Min: 10.55% Max: 85.64% Median: 39.965% Q1: 33.865 Q3: 46.675

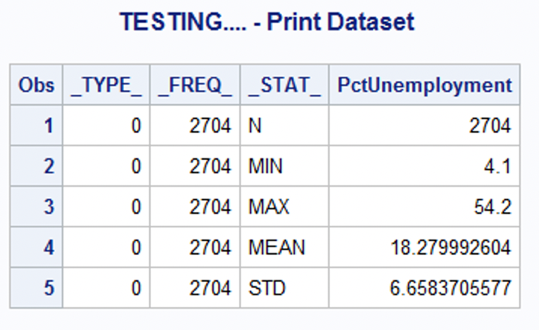
**MedianAge 5-Point Summary:**

 Table

Description automatically generated

MedianAge: Min: 20 Max: 55.4 Median: 34.3 Q1: 32.45 Q3: 36.4

**PctUnemployment 5-Point Summary:**

Table

Description automatically generated

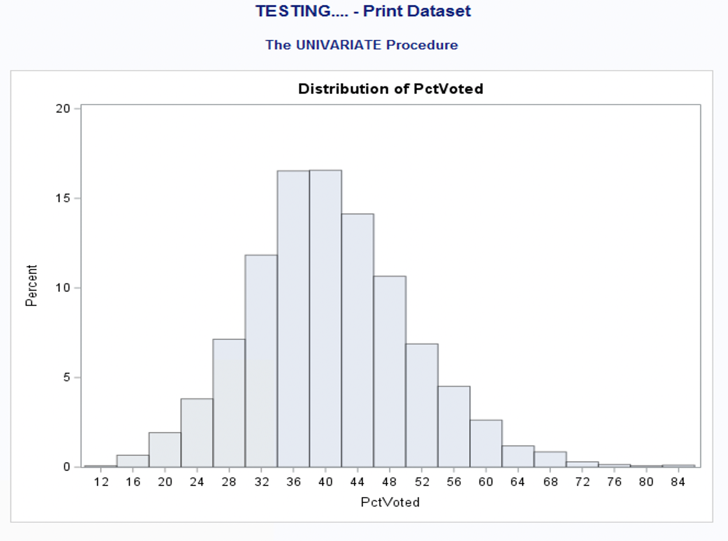
PctUnemployment: Min: 4.1% Max: 54.2% Median: 17.1%    Q1: 13.6%   Q3: 21.6%

Discussion:

Considering the data, out of 2704 observations, the minimum percent voter turnout by county was 10.55%, the minimum average voter age was 20 years old, and the minimum percent of unemployed voters was 4.1%. The maximum percent of voters by county was 85.64%, the average oldest/maximum voter(s) age was 55.4 years old, and the maximum percent of voters who are unemployed by county was 54.2%. The median voter turnout by county was about 39.97%, the median average age of voters was approximately 34.44 years old, and the median voters who were unemployed were 17.1%. Voters by county in the 25th percentile had 33.865% voter turnout, while those in the 75th percentile had 46.675% voter turnout. One standard deviation from the mean or average voter turnout (approx. 40.61%) is about +/- 10.18%. This means that, on average, counties had a 40.61% voter turnout with a standard deviation of 10.18%. Median Age voters by county in the 25th percentile were 32.45 years old, while those in the 75th percentile were 36.4 years old. One standard deviation from the mean or average median age (approx. 34.44 years old) is about +/- 3.53%. On average, people's ages of those who voted were 34.44 years old voter turnout with a standard deviation of 3.53 years old. Percent of voters by county who were unemployed in the 25th percentile was 13.6%, while those in the 75th percentile were 21.6%. One standard deviation from the mean or average voter turnout (approx. 18.28%) is about +/- 6.66%. On average, counties had about 18.28% of unemployed voters, with a standard deviation of 6.66%.

* e)  Create a histogram to analyze the percent of people who voted. Include the histogram output. Using the histogram and the 5-point summary from the previous question, analyze the histogram. Discuss your findings using the 5-point summary, normality (normal or skewed), modality, outliers, etc.?

**Histogram:**



5-Point Summary: Min: 10.55%   Max: 85.64% Median: 39.965%   Q1: 33.86%  Q3: 46.675%

Normality: Normal

Modality: Unimodal (One peak)

Outliers: There are no noticeable outliers in the histogram

Discussion:

Considering the histogram, we can tell our data is overall a normal bell curve with a single peak making it unimodal. Based on the chart, there don’t appear to be any noticeable outliers. Regarding the 5-point summary, we understand that there is a maximum observation of 85.64% voter turnout by county and a minimum of 10.55%. If we review the IQR, +/- 1.5 IQR (Q1 -1.5\*sd and Q3 + 1.5 \*sd), we can recognize that below Q1 -1.5\*sd (18.59%) and beyond such that Q3 + 1.5 \*sd (~61.95%) would be outliers. Therefore, outliers do exist in our data.

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

**Boxplot:**

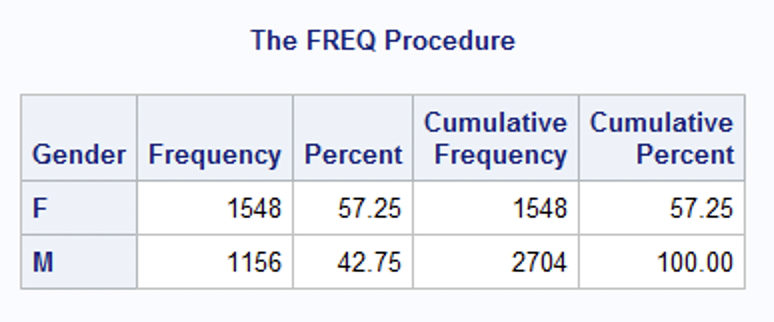
Chart, box and whisker chart

Description automatically generated

Discussion:

Considering the boxplot, we can tell that all female gendered observed outliers are + 1.5 IQR, therefore women are making up higher than typical voter percent rates in some counties. On the other hand, male gendered observed outliers exist +/- 1.5 IQR meaning that in some counties there are either lower than 1.5 IQR range of male voter percent rates or higher than 1.5 IQR. Overall, the data can be interpreted that a voter may be likelier a female than a male.

* g)  What is the gender breakdown in this dataset? (Hint: use PROC FREQ). Include the output. Which is the predominant gender in this dataset? Explain how you came to this conclusion.



Discussion:

The predominant gender is female. This is due to the greater frequency of females in comparison to males. As seen in the chart, those observed who were female made up 57.25% (1548/2704 of observations) of the cumulative percentage, while males make up the remainder at 42.75%(1156/2704 of observations.

* h)  Copy and paste your FULL SAS code into the word document along with your answers.

\*Import Statements;

**PROC** **IMPORT** file ="C:\Users\DVALDE12\Desktop\A1\election.txt"

out= elect replace;

**run**;

\*PROC PRINT data = elect (obs=5) noobs;

\*run;

\*PctVoted Summary;

**PROC** **SUMMARY** data = elect;

var PctVoted;

output out = summary\_PctVoted;

**run**;

**PROC** **PRINT** data = summary\_PctVoted;

\*PctVoted Median/Q1/Q3;

**PROC** **Means** data = elect Q1 Q3 Median;

var PctVoted;

**run**;

\*MedianAge Summary;

**PROC** **SUMMARY** data = elect;

var MedianAge;

output out = summary\_MedianAge;

**run**;

**PROC** **PRINT** data = summary\_MedianAge;

\*MedianAge Median/Q1/Q3;

**PROC** **Means** data = elect Q1 Q3 Median;

var MedianAge;

**run**;

\*PctUnemployment Summary;

**PROC** **SUMMARY** data = elect;

var PctUnemployment;

output out = summary\_PctUnemployment;

**run**;

**PROC** **PRINT** data = summary\_PctUnemployment;

\*PctUnemployment Median/Q1/Q3;

**PROC** **Means** data = elect Q1 Q3 Median;

var PctUnemployment;

**run**;

\*Create Histogram PctVoted;

**PROC** **UNIVARIATE** data = elect;

var PctVoted;

HISTOGRAM PctVoted;

**run**;

**PROC** **PRINT** data = PctVoted;

\*Create Boxplot PctVoted by Gender;

**PROC** **SGPLOT** data = elect;

vbox PctVoted / category = Gender;

**run**;

**PROC** **PRINT** data = elect;

\*Using PROC FREQ to determine Gender Breakdown;

**PROC** **FREQ** data = elect;

TABLE Gender;

**run**;