CSC423 – Assignment 1

**Q1.**

1. How many fields are there?
   1. There are a total of **7 Fields**. Namely: County, Pct\_Voted, MedianAge, MeanSavings, Pct\_Poverty, PopulationDensity and Gender.
2. Which Fields are numerical?
   1. The following fields are numerical: Pct\_Voted, MedianAge, MeanSavings, Pct\_Poverty and PopulationDensity.
3. Which fields are text?
   1. The following fields are text: County and Gender.

**Q2.** SAS code to create the SAS dataset using IMPORT is as follows:

title "Importing Dataset using IMPORT";

**proc** **import** datafile="voting\_1992.txt" out=voting\_1992 replace;

delimiter = '09'x;

getnames = yes;

datarow = **2**;

**proc** **print**;

**run**;

SAS code to create a dataset using INFILE is as follows:

title "Importing Dataset using INFILE";

**data** voting\_1992\_infile;

infile "voting\_1992.txt" delimiter='09'x missover firstobs=**2**;

input County $ Pct\_voted MedianAge MeanSavings Pct\_poverty PopulationDensity Gender $;

**proc** **print**;

**run**;

**Q3.** The first 5 rows of the output are as follows:

title "Importing Dataset using INFILE";

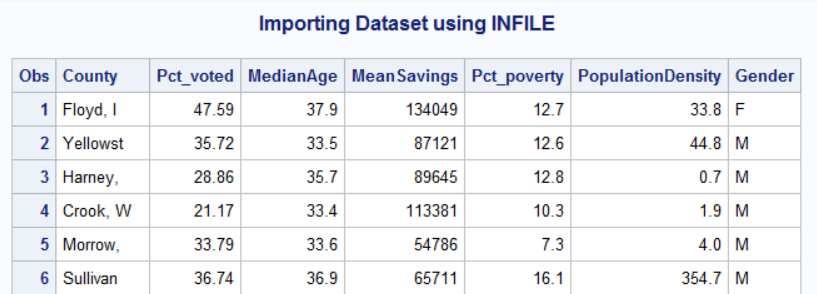
**data** voting\_1992\_infile;

infile "voting\_1992.txt" delimiter='09'x missover firstobs=**2**;

input County $ Pct\_voted MedianAge MeanSavings Pct\_poverty PopulationDensity Gender $;

**proc** **print**;

**run**;



**Q4.** The code for producing 5 point summary for People Voted and the median age is as follows:

title "5 Point Summary";

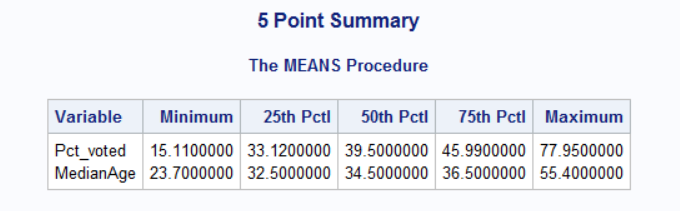
**proc** **means** min p25 p50 p75 max;

var Pct\_voted MedianAge;

**proc** **print**;

**run**;

The output of the above code is the following:



From the above table, we can make the following analysis about the Inter Quartile Range for Pct\_voted is 12.87 that means the outliers will lie outside the range of Q1 – 1.5\*IQR and Q3 + 1.5\*IQR which comes out to be (13.815,65.295) for Pct\_Voted. The range for Pct\_voted is 62.84.

IQR for MedianAge is 4 that means the range outside which any data point will be treated as an outlier is (26.5,42.5). The range for MedianAge is 31.70.

**Q5.** The code to create the histogram for the Pct\_voted attribute is as follows:

\*Histogram for the Pct\_voted;

title "Histogram with normal curve for Pct\_voted";

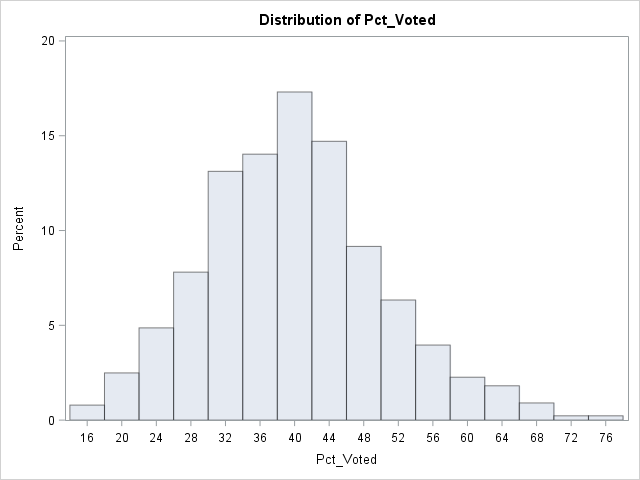
**proc** **univariate** normal;

var Pct\_voted;

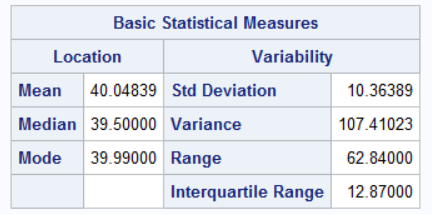
histogram Pct\_voted;

**run**;

The Histogram which is generated by the above code is as follows:



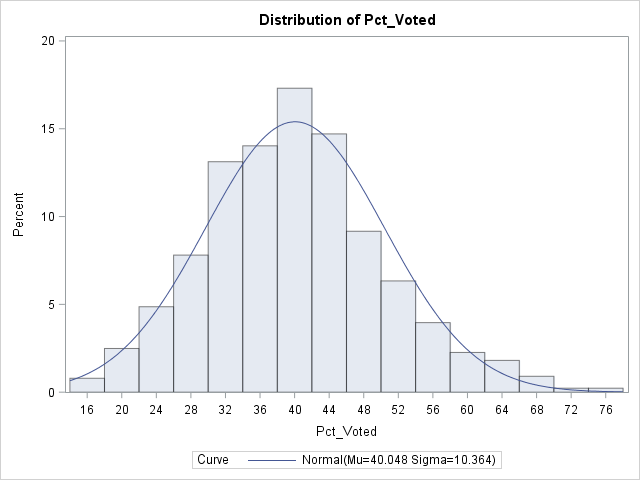
The basic statistical measures which were inferred using the histogram are as follows:



We can see that the median, range and the IQR are the same as we calculated from the 5-point summary. The Pct\_voted is normally distributed and to check whether it is so or not we can make use of SAS to superimpose a normal curve of the same attribute with the histogram.

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

By adding this piece of code snippet in the place of “histogram pct\_voted;” we get the following plot:



Hence, Pct\_voted is normally distributed.

As we had calculated in the previous question the values which lie outside the (13.815,65.295) will be considered as outliers. The maximum value for the Pct\_voted is 77.95 which is an outlier. Hence, there are outliers present. The median is the middle value and from the plot we can see that the Pct\_voted values near ~ 40% is the median which can be compared to the 5-point summary and the basic statistic measure table.

During the first observation of the above histogram we feel that the data is normally distributed but by observing the mean median and mode values we see that the mean > median which implies that the data is positively skewed. Though the difference is not huge and hence the skewness is also not much. The right tail also seems longer than the left tail which also indicates that the data is slightly positively skewed.

Median is 39.5 which also falls in the peak column in the histogram. Mean, Median Mode all lie in the same peak and so we can say that the skewness is not very large.

We can also say that in nearly 18% of the Counties Pct\_voted is around 38%-42% and the mode also lies in this range which can be inferred from the Basic Statistical Measure Table mentioned above.

**Q6.** The code to generate a boxplot between Pct\_voted and Gender is as follows:

\*Boxplot of Pct\_voted - Gender;

title "Boxplot of Pct-voted - Gender";

**proc** **sort**;

by Gender;

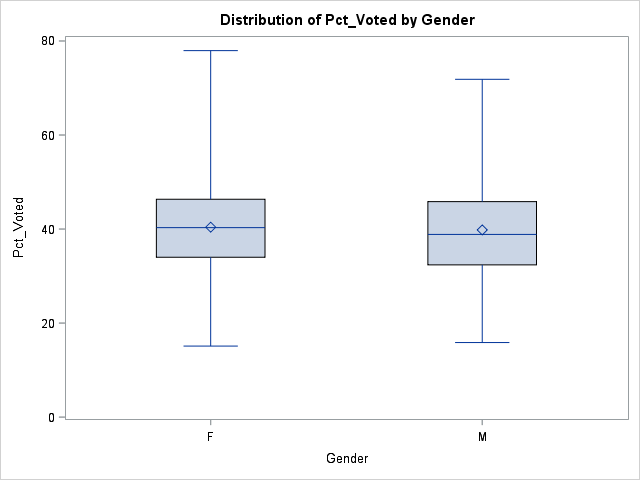
**run**;

**proc** **boxplot**;

plot Pct\_voted\*Gender;

**run**;

The box plot is as follows:



The things which we can infer from the boxplot are as follows:

* The median for the Pct\_voted in Female is around 40% and the median for the Male is also around 40%.
* The highest percentage of Female voters was around 79%-80% and for Male it was around 75%-78%.
* The lowest percentage of Male and Female voters is around 14%-16%.
* The Range for Pct\_voted in Females is (14%,79%) and for Male it is (14%,75%)
* The Median for the upper quartile of Pct\_voted for Male and Females seems to be the same at around 44%-46% (Pct\_voted).
* The Median for the lower quartile of Pct\_voted for Male and Females seems to be the same at around 34%-36%.
* From the boxplots we can also see that the median lies in the middle of the box, which implies the data does not sway ie it is normally distributed without much skew(skewness still exists).

**Q7.** The code for Gender Breakdown in the dataset is as follows:

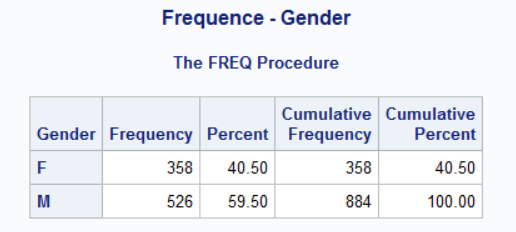
title "Frequency - Gender";

**proc** **freq**;

tables Gender;

**run**;

The table is as follows:



The predominant gender is Male(526) which is 59.5% of the total population (884).

The whole code for this assignment is as follows:

title "Importing Dataset using IMPORT";

**proc** **import** datafile="voting\_1992.txt" out=voting\_1992 replace;

delimiter = '09'x;

getnames = yes;

datarow = **2**;

**proc** **print**;

**run**;

title "5 Point Summary";

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

var Pct\_voted MedianAge;

**proc** **print**;

**run**;

\*Histogram for the Pct\_voted;

title "Histogram with normal curve for Pct\_voted";

**proc** **univariate** normal;

var Pct\_voted;

histogram Pct\_voted;

**run;**

\*Histogram for the Pct\_voted with a normal curve superimposed in the histogram;

title "Histogram with normal curve for Pct\_voted";

**proc** **univariate** normal;

var Pct\_voted;

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

**run;**

\*Boxplot of Pct\_voted - Gender;

title "Boxplot of Pct-voted - Gender";

**proc** **sort**;

by Gender;

**run**;

**proc** **boxplot**;

plot Pct\_voted\*Gender;

**run**;

title "Frequency - Gender";

**proc** **freq**;

tables Gender;

**run**;