**Data Analysis And Regression**

**Assignment-1** | **Total points: 10**

Note:

* All assignments should be submitted in a **single MS WORD format**, no PDFs or any other file types will be accepted. If you submit any other file type, it will not be graded.
* No extensions will be given unless for a documented reason specified in the syllabus, no late assignments past the due date even a couple of minutes late will be accepted as you have an extra day (8-days) to submit your assignments.
* Submitting work that is not yours is grounds for an automatic ‘F’ for the entire course – this includes taking content and ideas from others or consulting others to complete your deliverables other than your instructor.
* SAS software and virtual server stalls, gets slow and crashes; so start early and keep multiple backups in multiple places/mediums. Late submission or inability to do the assignment due to server and/or software issues will not be accepted. Any issues relating with SAS, contact IS using the phone number provided in the syllabus, I won’t be able to help you with DePaul software related issues.

**PROBLEM 1 [10 pts]**

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?

* There are 884 observations in the dataset.
  1. How many fields are there?
* There are 7 fields in the dataset
  1. Which fields are numerical?
* There are total 5 fields that are numerical. Which are **Pct\_Voted, MedianAge, MeanSavings, Pct\_Poverty,** and **PopulationDensity**.
  1. Which fields are text?
* There are total 2 fields that are text, **County**, and **Gender**.

1. 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.

\*Import Statement;

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

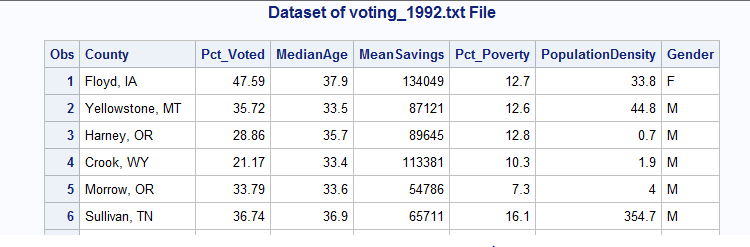
delimiter='09'x;

getnames=YES;

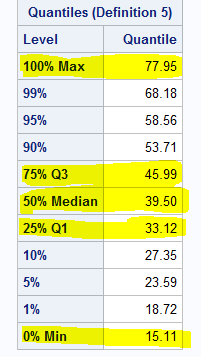
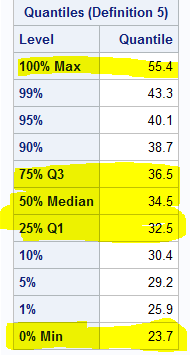
datarow=**2**;

**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.

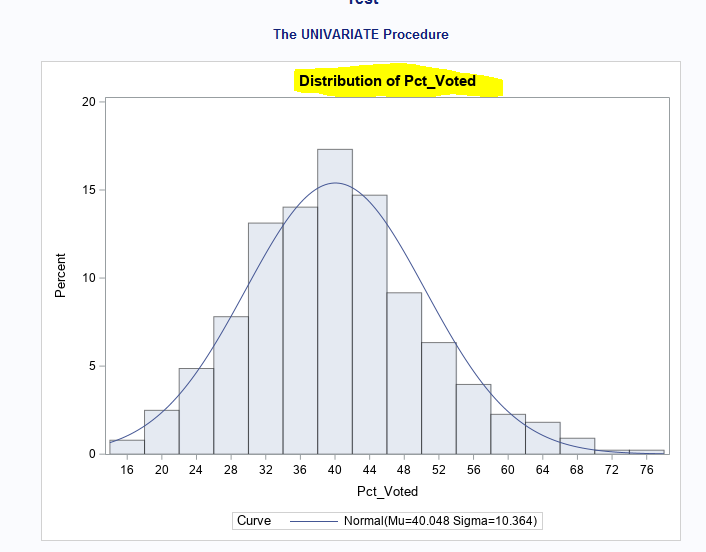
* 5num summaries for Pct\_voted and MedianAge are : -

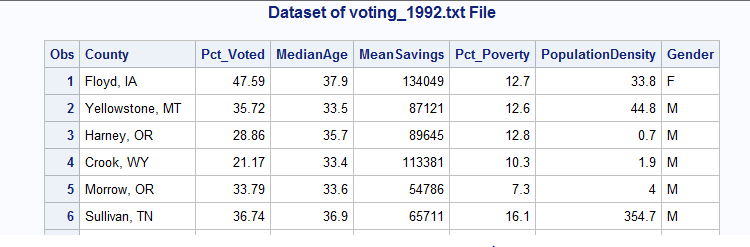
Pct\_Voted = 77.95, 45.99, 39.50, 33.12, 15.11

MedianAge = 55.4, 36.5, 34.5, 32.5, 23.7

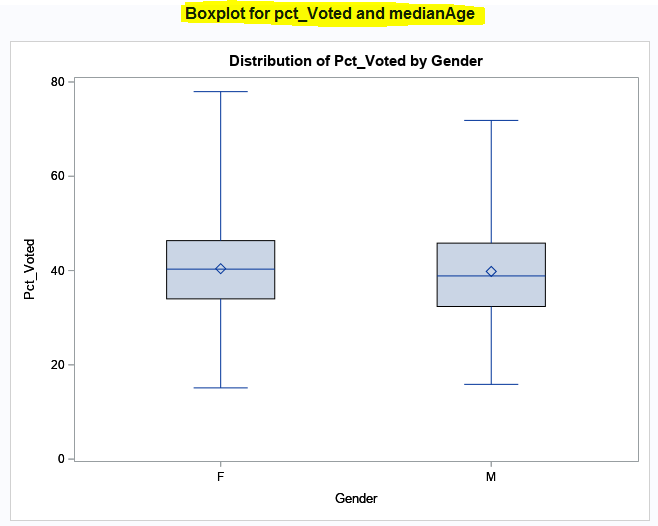
* Upon comparing both 5num summaries. We can see that as age went up, the percentage of people who voted also went up a drastic amount. The difference is huge when comparing min to the max but seems to be balanced when comparing the median, q1, and q3. So, we can assume that young people are not interested in voting, and their interest seems to increase as they age, and seems to increase more during their 30’s.

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 histogram appears to be normal, but its not perfectly normal as there is some frequency voted that are pulling the dataset a little towards the right. However, since those data points are not out of place, they does not have a drastic impact on the graph. The other thing that I noticed was that majority voted from 32-48 percent of the time. This number is telling us that over 50% of people are not voting.

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.

* ****
* After looking at the Boxplot we can say that Females tend to vote more when compared to male.

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

* The predominant gender in this dataset is Male being 526 (59.5%), and females being 358 (40.5%).

1. Copy and paste your FULL SAS code into the word document along with your answers. 3

\* TEST CODE - to test the data upload to server process is working;

\*Import Statement;

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

delimiter='09'x;

getnames=YES;

datarow=**2**;

**RUN**;

\*Print dataset voting;

TITLE 'Dataset of voting\_1992.txt File';

**PROC** **PRINT**;

**RUN**;

\* Print Histogram and 5Num Summary;

title "Histogram and 5Num Summary for pct\_Voted and medianAge";

**PROC** **univariate** DATA=voting normal;

var Pct\_Voted MedianAge;

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

**RUN**;

\*Print Gender Frequency;

title "Frequency for Gender";

**PROC** **FREQ** data=voting;

TABLE Gender;

**RUN**;

\* Print Boxplot for Pct\_voted and Gender;

title "Boxplot for pct\_Voted and medianAge";

**PROC** **SORT** data=voting;

BY Gender descending pct\_voted;

**PROC** **boxplot** data=voting;

plot Pct\_Voted\*Gender;

**RUN**;