**DSC423: Data Analysis And Regression / DSC 324: Data Analysis & Statistical Software II**

**Assignment-1** | **Total points: 15 for DSC 323 and DSC 423**

**Due Date: 04/09/2019 by 11:59 pm**

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

Examine the two code segments and answer the following questions.

***Code-1***

**data** cpu;

infile "cpudata.csv" delimiter=',' firstobs = 2;

input time line step device;

**run**;

***Code-2***

**proc** **import** datafile="cpudata2.txt" 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** cpudata.csv
2. **The datafile name used in Code-2 is** cpudata2.txt
3. **SAS dataset name for Code-1 is** cpu since it is not stated
4. **SAS dataset name for Code-2 is** cpu\_imp
5. **The delimiter used in Code-1 is (specify in words, do not copy and paste what’s given under delimiter) a comma**
6. **The delimiter used in Code-2 is (specify in words, do not copy and paste what’s given under delimiter) a blank space**
7. **The datafile extension of Code-1 is .csv**
8. **The datafile extension of Code-2 is .txt**
9. **Which line does the data start for Code-1? Firstobs=2 so line 2**
10. **Which line does the data start for Code-2?** DATAROW=**1 so Row 1**

**PROBLEM 2 [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? 7
   3. Which fields are numerical? Pct\_Voted, MedianAge, MeanSavings, Pct\_Poverty, PopulationDensity
   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.
3. 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.

| **Obs** | **County** | **Pct\_Voted** | **MedianAge** | **MeanSavings** | **Pct\_Poverty** | **PopulationDensity** | **Gender** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Floyd, IA | 47.59 | 37.9 | 134049 | 12.7 | 33.8 | F |
| **2** | Yellowstone, MT | 35.72 | 33.5 | 87121 | 12.6 | 44.8 | M |
| **3** | Harney, OR | 28.86 | 35.7 | 89645 | 12.8 | 0.7 | M |
| **4** | Crook, WY | 21.17 | 33.4 | 113381 | 10.3 | 1.9 | M |
| **5** | Morrow, OR | 33.79 | 33.6 | 54786 | 7.3 | 4 | M |

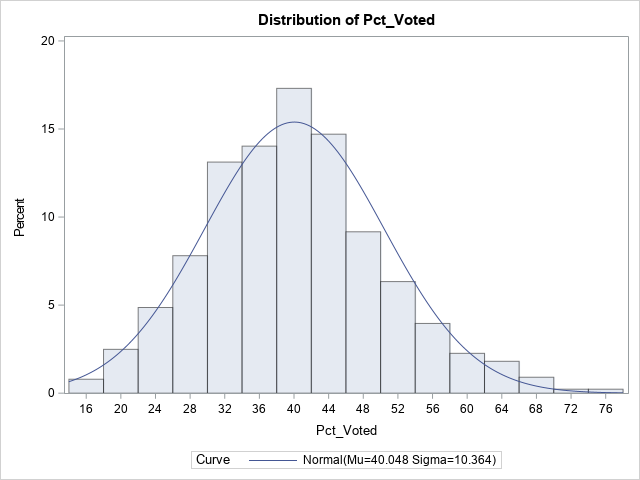
1. What is the 5-point summary numbers for percentage of people voted and median age? The 5-point summary numbers are min, max, median or 50% percentile, Q1 and Q3. Include the output. Discuss your findings.

The MEANS Procedure

| **Variable** | **Mean** | **Minimum** | **25th Pctl** | **50th Pctl** | **75th Pctl** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Pct\_Voted** | | **MedianAge** | | |  | | --- | | 40.0483937 | | 34.4925339 | | |  | | --- | | 15.1100000 | | 23.7000000 | | |  | | --- | | 33.1200000 | | 32.5000000 | | |  | | --- | | 39.5000000 | | 34.5000000 | | |  | | --- | | 45.9900000 | | 36.5000000 | | |  | | --- | | 77.9500000 | | 55.4000000 | |

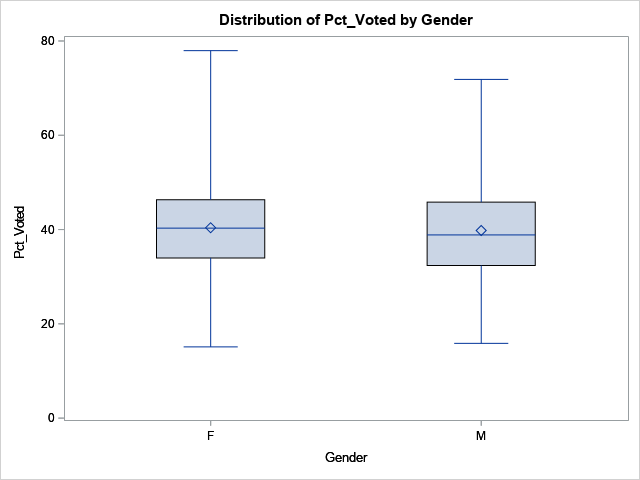
The mean % of people that voted in all counties was 40.048393% and the median age of voters was 34.4925339. Less than half of the population in all county’s was voting and the voters were primarily younger in age. The minimum shows that one county only had 15% of people that voted and the max was 77.95% in another county. The lowest age voter was 23.7 years (assuming voting age starts at 18 years old, 23.7 is higher than expected for the minimum. The oldest person being 55.4 that voted is also younger than expected.) The inner 50% gives the most info and disparity of the data. For percent voted, the range for inner 50% is 33.12 and 45.99 percent, showing that this does not even cross the halfway mark for the percentage of voters in a county. Median age being 32.5 to 36.5 is young, but seeing that the max age was 55.4, this age range falls in a higher range since voting is in a field of 18-55.4 years of age.

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?



Because percentage goes from 0-100, a value of 15 or 77.95 would not be considered outliers because they are a part of the normal percentage distribution. In the histogram, Mu is listed as 40.084 and this matches up with the descriptive that was given for mean in the previous problem. The distribution is normal and not skewed left or right.

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 percent of people that voted is on the y axis and gender is on the x axis. The plot on the left shows female analytics and the right shows male. The boxplots seem to be very similar at first glance, however, the max percentage of female voters in a county was higher than the max of males.

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

| **Gender** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **F** | 358 | 40.50 | 358 | 40.50 |
| **M** | 526 | 59.50 | 884 | 100.00 |

The predominant gender is Males by almost 200.

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

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

delimiter = '09'x;

getnames = yes;

datarow = **2**;

**run**;

title "print dataset using import";

**proc** **print**;

**run**;

title "descriptives";

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

var Pct\_Voted MedianAge;

**run**;

title "histogram";

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

var Pct\_voted;

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

**run**;

title "boxplot";

**proc** **sort**;

by gender;

**run**;

**proc** **boxplot**;

pot pct\_voted\*gender; \*y-axis = pt\_voted and x-axis = gender;

**run**;

title "frequency of gender";

**proc** **freq**;

table gender;

**run**;