**Ximan Liu**

**Data Analysis And Regression**

**Assignment-1** | **Total points: 15**

**Due Date: 04/15/2021 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 \_\_\_\_\_\_\_\_\_**
2. **The datafile name used in Code-2 is \_\_\_\_ cpudata2\_\_\_\_\_\_\_\_\_\_**
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 (specify in words, do not copy and paste what’s given under delimiter) \_\_\_\_\_comma\_\_\_\_\_\_\_\_\_\_**
6. **The delimiter used in Code-2 is (specify in words, do not copy and paste what’s given under delimiter) \_\_\_\_\_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? \_\_\_\_2\_\_\_\_\_\_\_\_\_\_\_\_\_\_**
10. **Which line does the data start for Code-2? \_\_\_\_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**

* 1. How many fields are there?

**7**

* 1. Which fields are numerical?

**Pct\_Voted, MedianAge, MeanSavings, Pct\_Poverty, PopulationDensity**

* 1. Which fields are text?

**County, 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.

**TITLE "Census Records Selected Counties in the U.S. Who Voted in 1992 Elections";**

**PROC IMPORT datafile = "C:\Users\XLIU115\Desktop\Assignment1\voting\_1992.txt" out = county 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.

Table

Description automatically generated

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.

Table

Description automatically generated

**Pct\_Voted:**

**The median for percentage of people voted is 39.5%, which is quite low in general. The voting situation in some counties shows lack of participation which only 15.11% of people voted. Vice versa, in some counties the voting situation is quite active, which 77.95% of people get involved in voting. The median 50% percentile is between 33.12% to 45.99%, in other words it is located between Q1 and Q3.**

**MedianAge:**

**The median for percentage of people voted is 34.5. Some voted people show that their ages are quite young, which starting from 23.7-year-old. Vice versa, the elder voted people after 55.4-year-old tend not to vote, which indicates the oldest age of voters. The median age is located in the middle 50% percentile.**

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?

Chart, histogram

Description automatically generated

**The histogram distribution of percentage people voted is normal distribution with a bell-shaped density curve. The median 39.5% is quite close to the mean 40.04839%. The graph is symmetric and unimodal with skewness 0.392301. The highest number of percentage voted people is 40%. From the data coming with the histogram, we can tell that Q1 is 33.12%, Q3 is 45.99%, and IQR is 12.87%. To compute outliers, 33.12% - 12.87% \* 1.5 = 13.815% is the low outlier; Vice versa, 45.99% + 12.87% \* 1.5 = 65.295% is the high outlier. In other words, numbers out of 13.815% to 65.295% are outliers.**

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.

Chart, box and whisker chart

Description automatically generated

**The medians of female and male majority gender are quite close. Upper quartile (Q3) of percentage of voted people for both genders are lower than 50%. Female has longer range of distribution in the graph, from lower extreme 16% to upper extreme 79%. For both genders, the medians are quite close to the means. Precisely speaking, the median and mean of female are more overlapped. There is no outlier showed in the graph.**

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

Table

Description automatically generated

**For female, the gender breakdown is 40.5% and 59.5% for male. Male is the majority gender of voters. Female is majority gender in 358 counties. Male is majority gender in 526 counties.**

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

**\*2B;**

**TITLE "Census Records Selected Counties in the U.S. Who Voted in 1992 Elections";**

**PROC IMPORT datafile = "C:\Users\XLIU115\Desktop\Assignment1\voting\_1992.txt" out = county replace;**

**delimiter = '09'x;**

**getnames = yes;**

**datarow = 2;**

**RUN;**

**\*2C;**

**PROC print;**

**RUN;**

**\*2D;**

**TITLE "Descriptives";**

**PROC MEANS data = county min max median p25 p75;**

**VAR Pct\_Voted MedianAge;**

**RUN;**

**\*2E;**

**TITLE "Histogram";**

**PROC UNIVARIATE normal;**

**VAR Pct\_Voted;**

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

**inset median mean range kurtosis skewness;**

**RUN;**

**\*2F;**

**DATA county;**

**set county;**

**length Majority\_gender $10;**

**if gender='M' then Majority\_gender='Male';**

**else Majority\_gender='Female';**

**RUN;**

**TITLE "Boxplot - Percentage of People Voted by Gender";**

**PROC SORT;**

**by Majority\_gender;**

**RUN;**

**PROC boxplot;**

**plot Pct\_Voted\*Majority\_gender;**

**RUN;**

**\*2G;**

**TITLE "Gender Breakdown";**

**proc freq;**

**tables Majority\_gender;;**

**RUN;**