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

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.

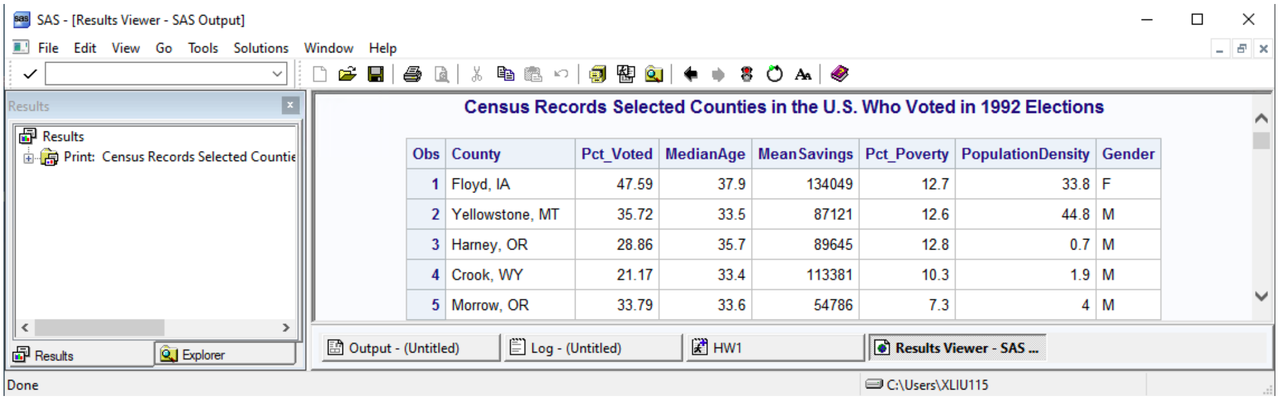
a) Open the dataset and examine the data. Answer the following:

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

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.

```
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;
```

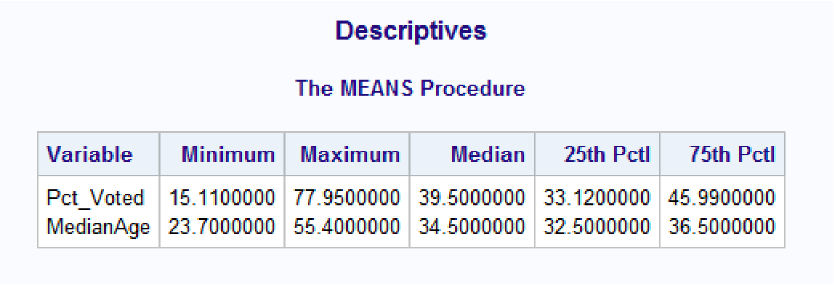
c) Run a PROC PRINT to print your dataset in SAS. Do a print screen, to copy and paste the first 5 observations of the output.


SAS - [Results Viewer - SAS Output]
File Edit View Go Tools Solutions Window Help
Results
Results
Print: Census Records Selected Countie
Results
Explorer
Output - (Untitled)
Log - (Untitled)
HW1
Results Viewer - SAS ...
Done
C:\Users\XLIU115

Census Records Selected Counties in the U.S. Who Voted in 1992 Elections

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

d) 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.



Descriptives

The MEANS Procedure

Variable	Minimum	Maximum	Median	25th Pctl	75th Pctl
Pct_Voted	15.1100000	77.9500000	39.5000000	33.1200000	45.9900000
MedianAge	23.7000000	55.4000000	34.5000000	32.5000000	36.5000000

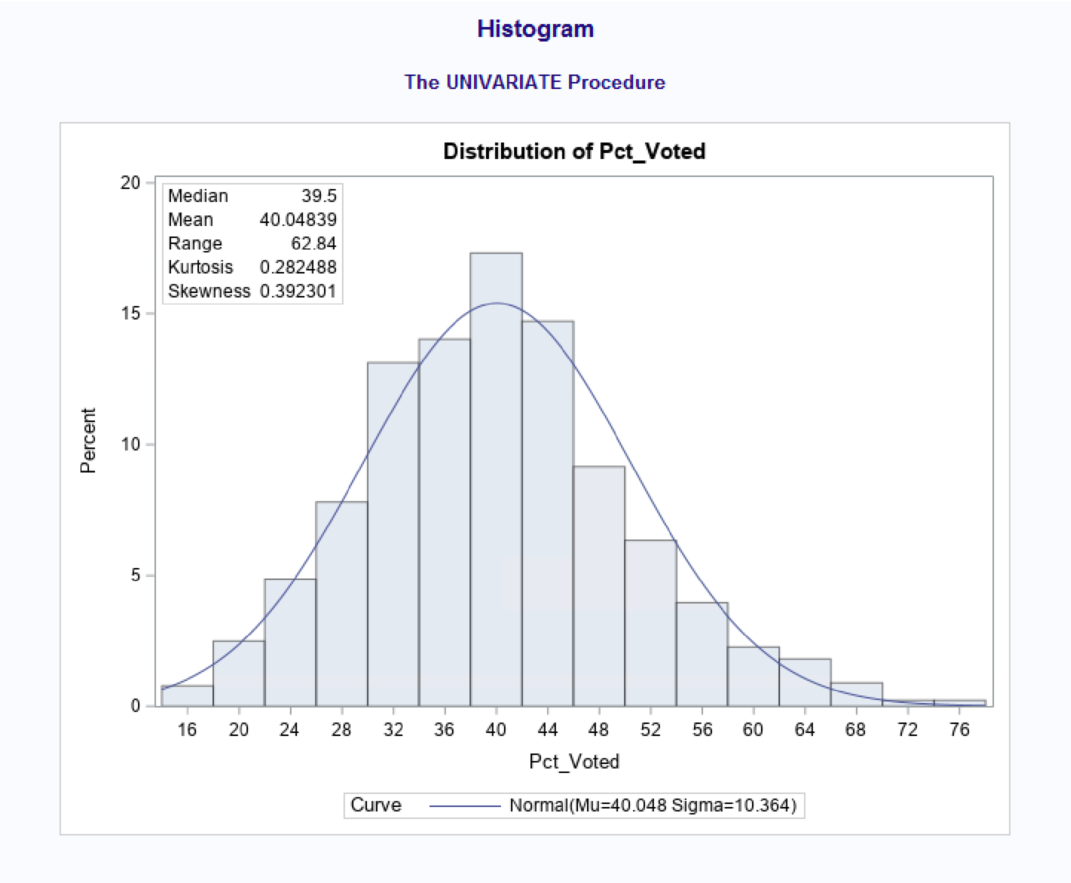
Pct_Voted:

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.

e) 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?



Histogram

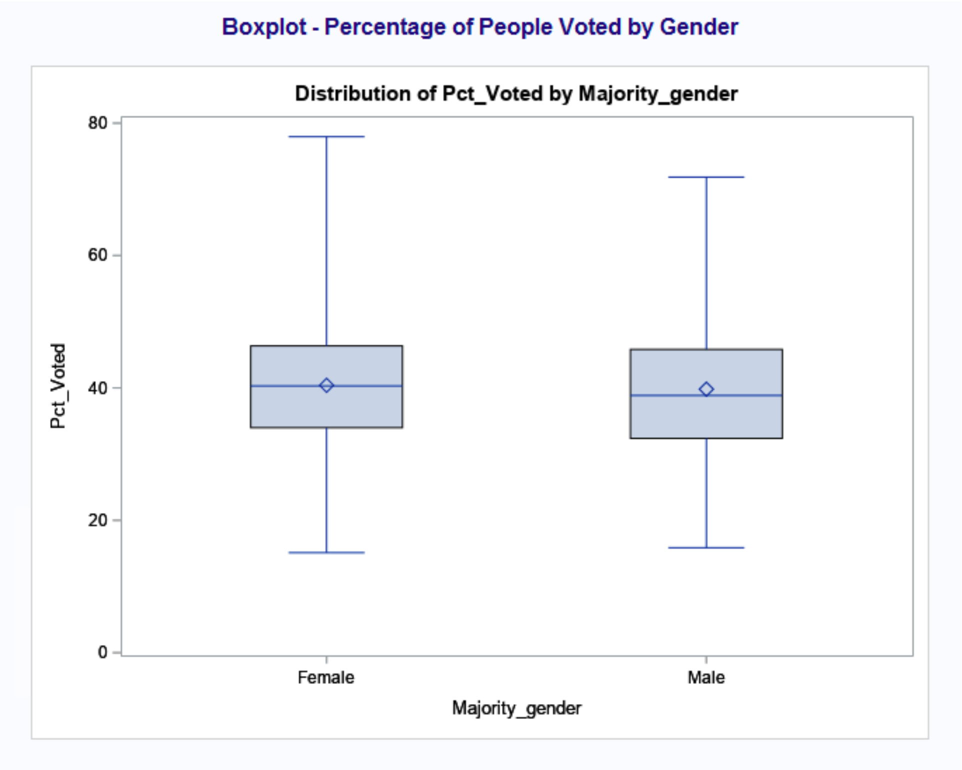
The UNIVARIATE Procedure

Distribution of Pct_Voted

Median	39.5
Mean	40.04839
Range	62.84
Kurtosis	0.282488
Skewness	0.392301

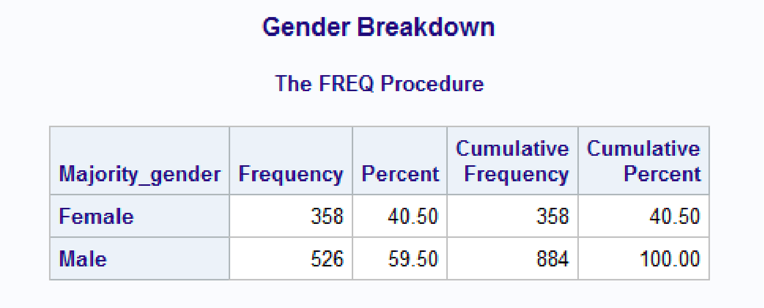
The histogram distribution of percentage people voted is normal distribution with a bell-shaped density curve. The median 39.5% is quite close to the mean 40.04839%. The graph is symmetric and unimodal with skewness 0.392301. The highest number of percentage voted people is 40%. From the data coming 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.

f) 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 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.

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



Gender Breakdown

The FREQ Procedure

Majority_gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	358	40.50	358	40.50
Male	526	59.50	884	100.00

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.

h) 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;
```