**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.**
   2. How many fields are there? **There are 7 fields**
   3. Which fields are numerical? **Numerical fields- Pct\_Voted, MedianAge, MeanSavings, Pct\_Poverty, PopulationDensity.**
   4. Which fields are text? **Text fields-County and 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.

**First 5 observations of Voting 1992 from proc print:**

![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]()

**My findings Percent of people voted:** The max county that voted was 77.9%. That number does not surprise me because only citizens can vote not everyone is a citizen of the U.S. why it is not near 100% percent voted. Another reason why the maximum is closer towards 75% because the choices on who to vote for in the election might not align with people’s views so they chose to skip. Voting days usually tend to be on weekdays and people work and it does not fit into their schedule. Why I see why the minimum is at 15% in low income areas I would assume people were not as aware or told about campaigning for elected officials so they choose it is not important for them to use their right to vote. The Q1, median, and Q3 make me aware that as a country in whole campaigning for elections and to vote is not as big deal as 2020 was just blasted across every platform. Low turnout of the 884 recorded observations means political interest across the country was not big then.

**My findings Median Age:** Seeing that the maximum for median age was only 55 surprised me that voting seems to me at least that it would be related to more older communities meaning the maximum in some counties would be higher. The Q1, median, and Q3 were within 13 years of each other ranging from 23-36 ½ meaning that younger adults and young parents in these counties were more willing to go out and vote. That could be a reason why in 1992 making elders go out and voting in person booths maybe was a factor why they did not come out compared to being able to mail vote. I think mail vote when it came out is able to get more people of wider range of ages to vote.

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]()

**My findings on the histogram:** This histogram output is a normal curve. From my understanding there are no outliers from this graph because this graph rises steadily then hits the peak of the data smoothly then consistency continues to go down the further the data goes. There is no outliers visible because there is no gaps between points of plotted data on this histogram. This histogram throughout is unimodal with only one peak.

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]()

**My findings on the boxplot**: Looking at the gender differences when it came to vote in 1992. Females maximum came out to vote more than the males. The range was around 10% higher maximum with right under 80% than males ~70% maximum turn out. The Q3 and Q1 for females are both higher than males. Q3 for females close to 50% while males Q3 under 50%. Q1 for females was ~35% while males Q1 was ~30%. The median for females was higher than males; it was rather close within about 1%-2% females around 40% and males at 38%. The only number summary that was higher for males was their minimum voting was slightly higher at ~18% than females ~15%.

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]()

**My findings on frequency**: Looking at the frequency of gender there are more males in this voting in this election. Males had a percent of 59.50% where females had 40.50%. Males had a 19% higher percentage to vote than females. Males frequency to vote was 526 which is more than females with 358. That number comes out to a 168-person difference between males and females.

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

\* PROC IMPORT;

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

DELIMITER='09'x;

GETNAMES=yes;

DATAROW=2;

RUN;

TITLE"Voting 1992";

PROC PRINT;

RUN;

\*Descriptives;

TITLE"Descriptives";

PROC MEANS max min p25 p50 p75;

VAR Pct\_Voted MedianAge MeanSavings Pct\_Poverty PopulationDensity;

RUN;

\*Histogram;

PROC UNIVARIATE normal;

VAR Pct\_Voted;

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

RUN;

\*Sort by Gender;

PROC SORT;

BY Gender;

RUN;

\*Boxplot;

Title"Percentage voted by Gender";

PROC Boxplot;

plot Pct\_Voted\*Gender;

RUN;

\*Frequency;

TITLE"Frequency - Gender";

PROC FREQ;

TABLES Gender;

RUN;