**CSC423: Data Analysis And Regression / CSC 324: Data Analysis & Statistical Software II**

**Summer Assignment-1** **CSC 423**

**Kristen Groom**

**Online student**

**7/23/2018**

**PROBLEM 1 [5 pts] – to be answered by everyone**

Examine the 2 code segments and answer the following questions.

***Code-1***

**data** cpu;

infile "cpudat.txt" delimiter=',';

input time line step device;

**run**;

***Code-2***

**proc** **import** datafile="cpudat2.csv" 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 :** *cpudat.txt*
2. **The datafile name used in Code-2 is :** *cpudat2.csv*
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:** *a comma - ‘,’*
6. **The delimiter used in Code-2 is:** *a space – ‘ ‘*
7. **The datafile extension of Code-1 is:** *.txt*
8. **The datafile extension of Code-2 is :** *.csv*
9. **Tue or False? “Code-1 uses import statement while Code-2 uses infile statement”:** *False*
10. **Tue or False? “Code-1 and Code-2 perform the same type of function – i.e. they both allow you to get the data written into SAS dataset”:** *True*

**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?
   2. How many fields are there?
   3. Which fields are numerical?
   4. Which fields are text?

*Answer:*

1. *How many Observations are there?*

*There are 884 Observations in the voting dataset*

1. *How many fields are there?*

*There are 7 fields or variables*

1. *Which fields are numerical?*

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

1. *Which fields 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.

*Answer:*

*TITLE "Get voting data into SAS Dataset";*

*PROC IMPORT datafile='S:\CSC423\HW1\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.

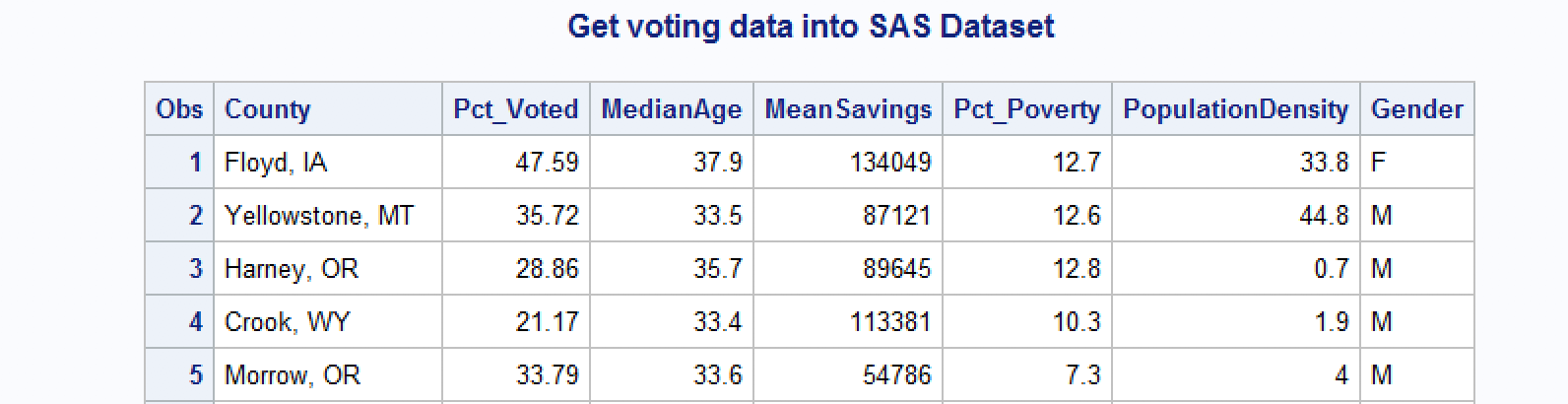
*Answer:*

***Code****:*

*PROC PRINT data=voting;*

*RUN;*

***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.

*Answer:*

***Code****:*

*\*Descriptives;*

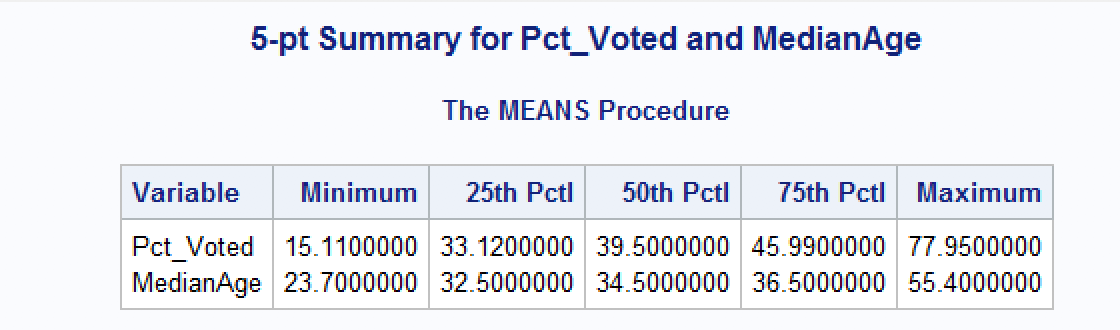
*TITLE '5-pt Summary for Pct\_Voted and MedianAge';*

*PROC means min p25 p50 p75 max;*

*var Pct\_Voted MedianAge;*

*RUN;*

***Output:***

******

***Findings:***

*I can see from the output that the minimum percentage of people who voted in a county was approximately 15% and the maximum was approximately 78%. I can see that a quarter of all the data for percentage of voters fall below approximately 33%, that 50% of all counties had approximately up to 39.5% voters and 50% had more than 39.5% voters. I see that three-quarters of the counties had approximately 46% voters.*

*I also see from looking at the median age descriptive statistics that the lowest median age for counties is approximately 24 years, while the oldest median age for a county is approximately 55 years. I can see that 75% of all data for median age fall below approximately 37 years, 50 percent of the data is below approximately 35 years of age, and that a quarter of the data represents a median of approximately 33 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?

*Answer:*

***Code:***

*\*Histogram and normal curve for Percentage of voters;*

*TITLE 'Histogram + normal curve for Pct\_Voted';*

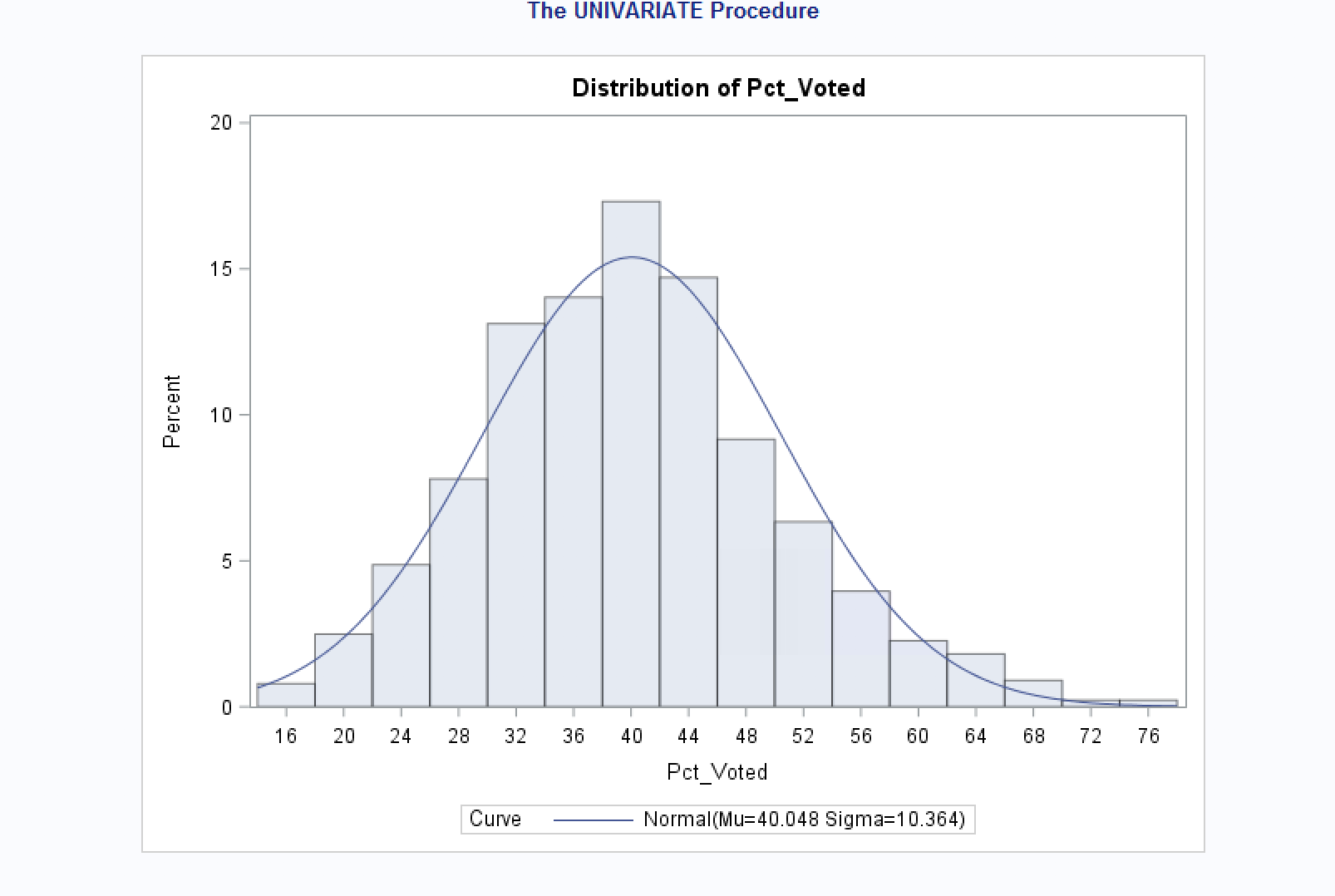
*proc univariate normal;*

*var Pct\_Voted;*

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

*run;*

***Output:***



***Findings:***

*The histogram and the normal curve shows us that the data is almost normally distributed but slightly positively skewed or skewed to the right. The histogram confirms our findings from the 5-point summary for percentage of voters in that we see that the median is at approximately 40% with fairly symmetric tails showing 25% of data around and below approximately 33% and 75% of the data falling around and below 46%. I do not see outliers but this is not completely clear from the data. We would have to perform further statistics to confirm whether there are actually outliers at the 75%+ percentage of voters.*

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.

*Answer:*

***Code:***

*\*Boxplot;*

*TITLE 'Boxplot showing percentage of voters by gender';*

*\*Sort by the gender column by-which -> gender;*

*PROC SORT;*

*by Gender;*

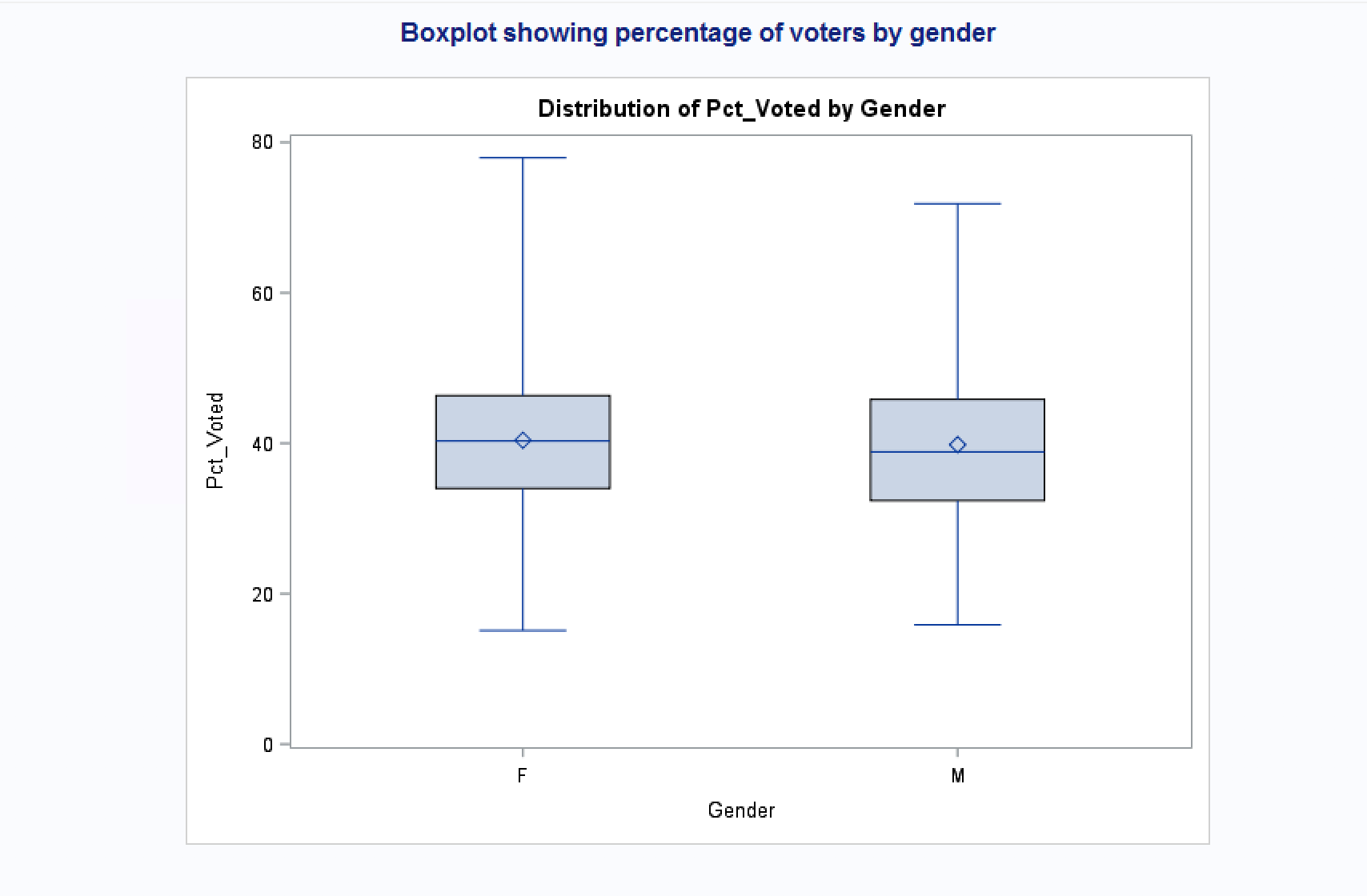
*run;*

*proc boxplot;*

*plot Pct\_Voted\*Gender;*

*run;*

***Output:***

******

***Findings:***

*Looking at the boxplot, I see that the median for dominant gender in our observed counties ie. males and females of the people who voted are almost equal with both being around 40%, although it seems males as the dominant gender have a slightly lower percentage of voter turnout. The range for female voters is larger and we can see the maximum percentage of voters as predominantly females at almost 80% while males appear to be closer to 75%. Both males and females have minimum percentage of voters that are approximately 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?

*Answer:*

***Code:***

*\*Frequency table for Gender;*

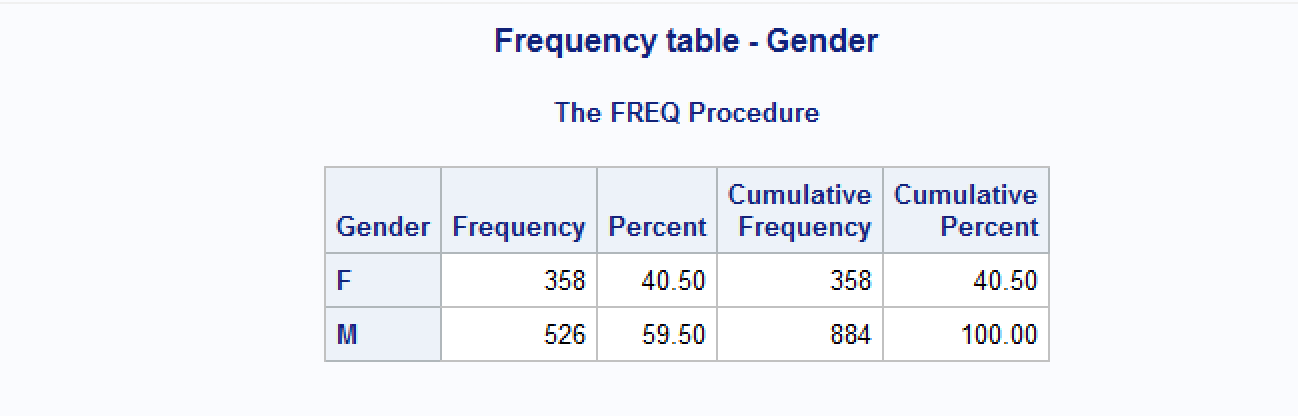
*TITLE 'Frequency table - Gender';*

*PROC freq;*

*tables Gender;*

*run;*

***Output:***

******

***Findings:***

*We can see that the data does not equally represent predominantly male and female represented counties which could explain some of the differences we see in the boxplot for Gender. Only 40.50% of our data identify as Female, and 59.50% of our data observations represent those who identify as Males.*

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

*Answer:*

*This has been done above.*

**The goal of problem 3 is to apply linear regression to a dataset using a statistical package.**

***Note: For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.***

**PROBLEM 3 [15 pts] – to be answered by everyone**

The file banking.txt attached to this assignment provides data acquired from banking and census records for different zip codes in the bank’s current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices. The data show

* median age of the population (AGE)
* median income (INCOME) in $
* average bank balance (BALANCE) in $
* median years of education (EDUCATION)

In this exercise you are asked to apply regression analysis techniques to describe the effect of age education and income on average account balance.

***Code to input dataset:***

*\* Get voting.txt into dataset*

*TITLE 'Banking Dataset using INFILE';*

*DATA banking;*

*INFILE 'S:\CSC423\HW1\Banking.txt' expandtabs missover firstobs=2;*

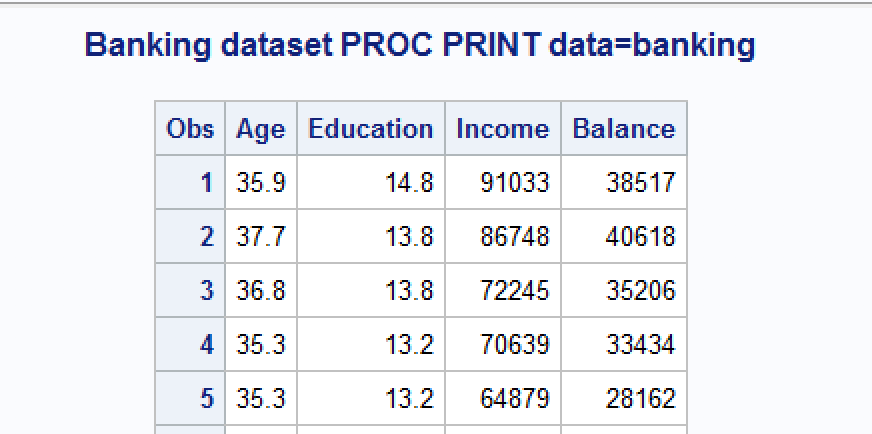
*INPUT Age Education Income Balance;*

*run;*

*PROC PRINT data=banking;*

*RUN;*

***Dataset voting output (head of dataset):***

******

1. Analyze the distribution of average account balance using histogram, and compute appropriate descriptive statistics. Write a paragraph describing distribution of Balance and use appropriate descriptive statistics to describe center and spread of the distribution. Discuss your findings. Also, do you see any outliers? Include the histogram.

*Answer:*

***Code:***

*\*Histogram and normal curve for Account Balance;*

*TITLE 'Histogram + normal curve for Balance';*

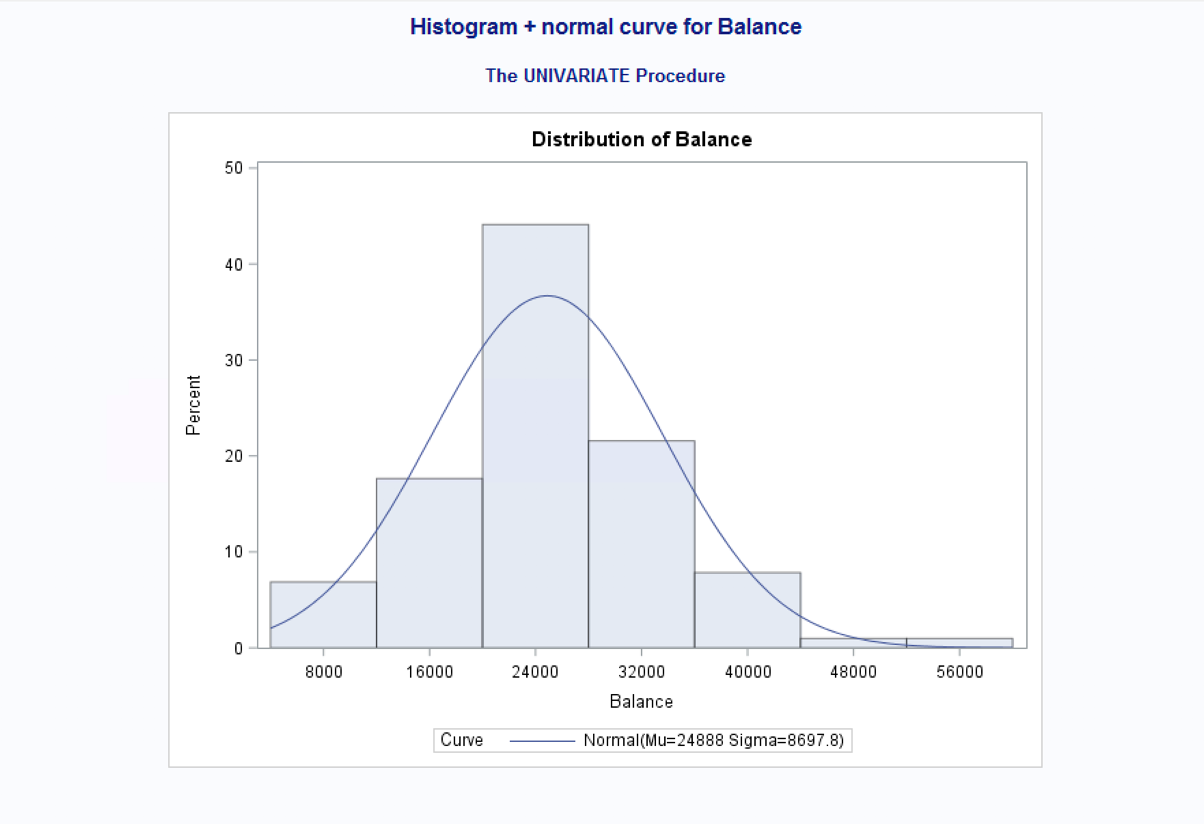
*proc univariate normal;*

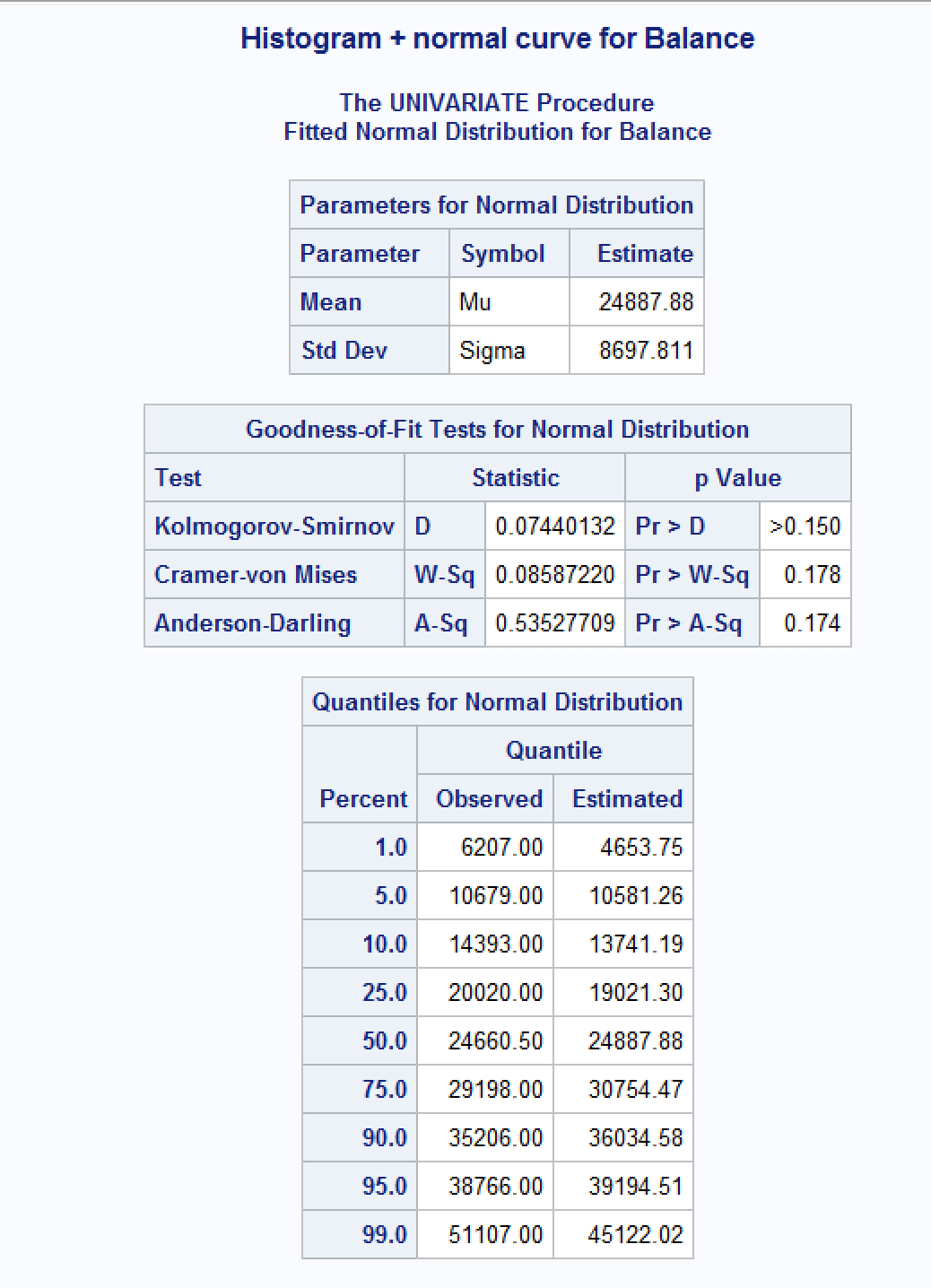
*var Balance;*

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

*run;*

***Output:***

******

******

***Findings:***

*In looking at the histogram output, I see a fairly normally distributed plot for the Account Balance data. It appears to have a slight right tail so it has a slightly positively skewed look to it but only slightly and I can verify this by looking at the Mean and Median which are close in amount with the mean being $24887.88 which is only approximately $200 more than the Median which is $24660. This tells me that there may be a pull to the right with possible outliers. I can see that 95% of the balances will be less than or equal to $38,766.00 and then a large increase in the balances after this 95th percentile, so there is at least one possible outlier for the balance above $51,00. The standard deviation for Balance is $8697.811 which tells me that approximately 68% of balances will be between and approximately $16,190.07 and $33,584.88, and that approximately 95% of all balances will be between the approximate values of $7,492.26 and 42,283.502. It also tells us that approximately 99% of balances will fall within the range of <0 to approximately $50,981.31. This also suggests the possibility of $51107.00 being an outlier.*

1. Create scatterplots to visualize the associations between bank balance and the other variables. Discuss the patterns displayed by the scatterplot. Also, do the associations appear to be linear? (You can create scatterplots or a matrix plot). Include the scatterplots.

*Answer:*

***Code:***

*\*Gplot for balance and x vars;*

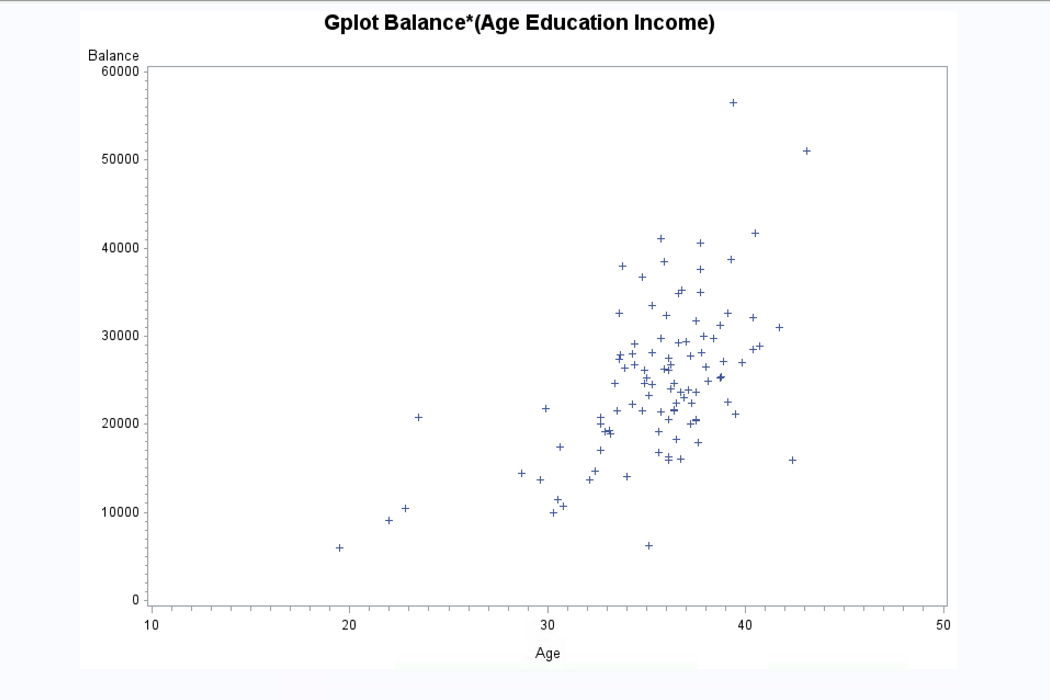
*TITLE 'Gplot Balance\*(Age Education Income)';*

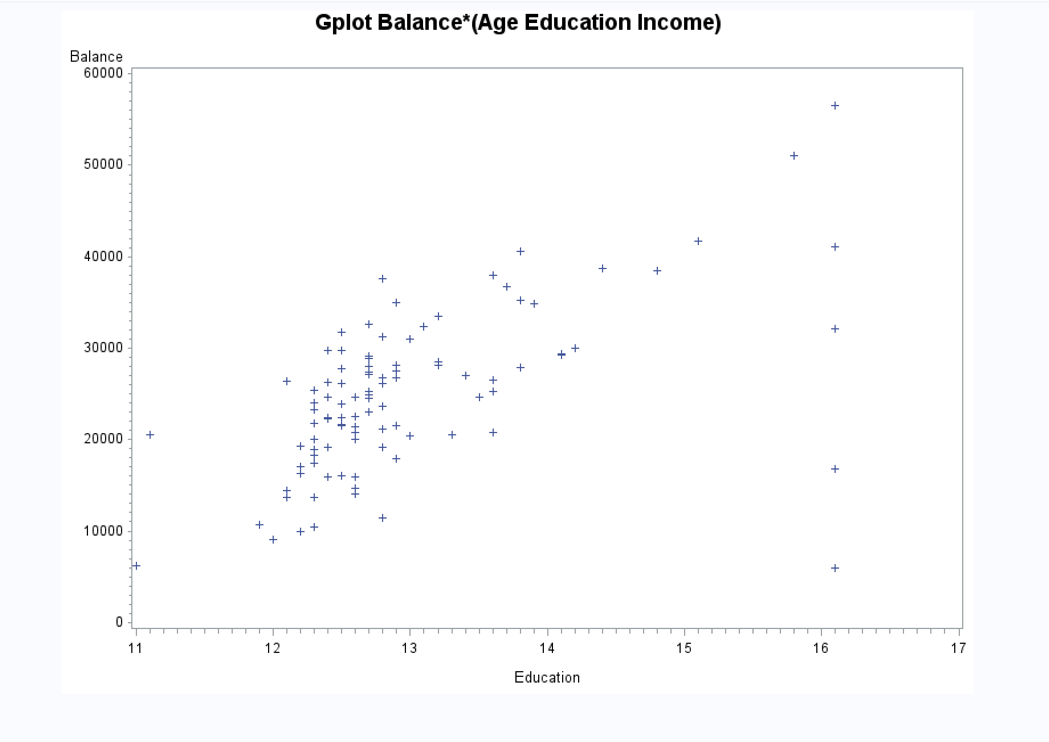
*PROC GPLOT;*

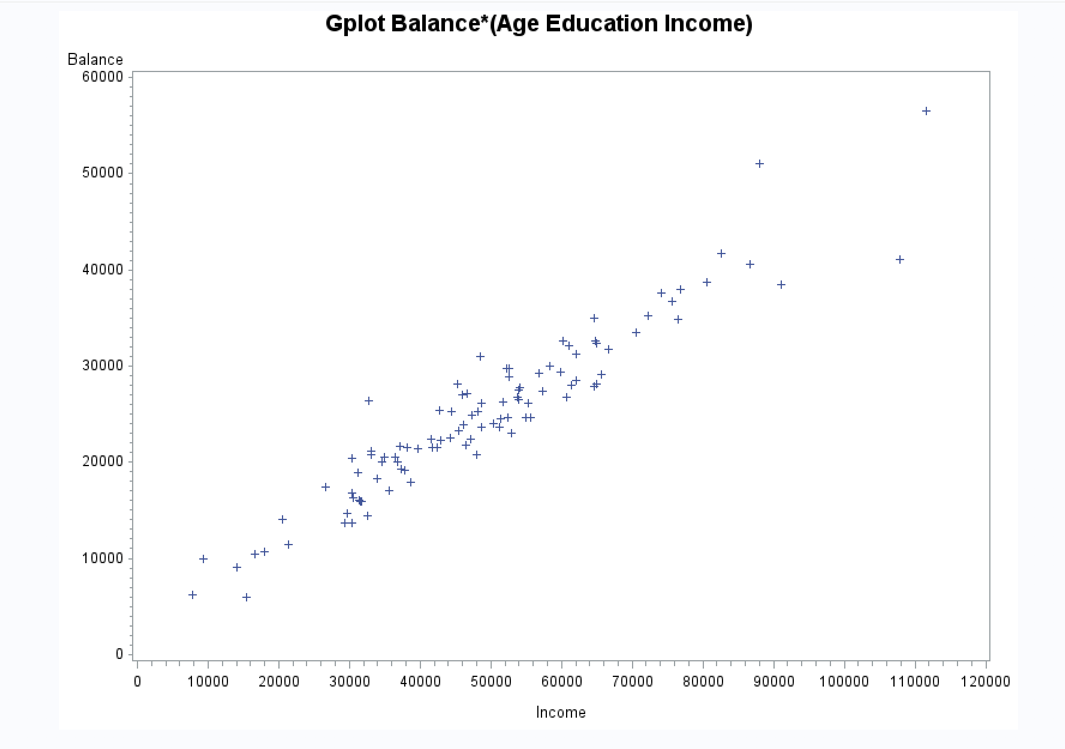
*PLOT Balance\*(Age Education Income);*

*RUN;*

***Output:***

**

**

**

***Findings:***

*I see linear associations between balance and each of the predictor variables of Age, Education, and Income. They all appear to have a positive association and all three seem somewhat strong or moderate but I would have to view the correlation values to find how strong the association actually is.*

1. Compute correlation values of bank balance vs the other variables. Interpret the correlation values, and discuss which pairs of variables appear to be strongly associated. Include the relevant output that shows the correlation values.

*Answer:*

***Code:***

*\*Correlation Analysis for banking;*

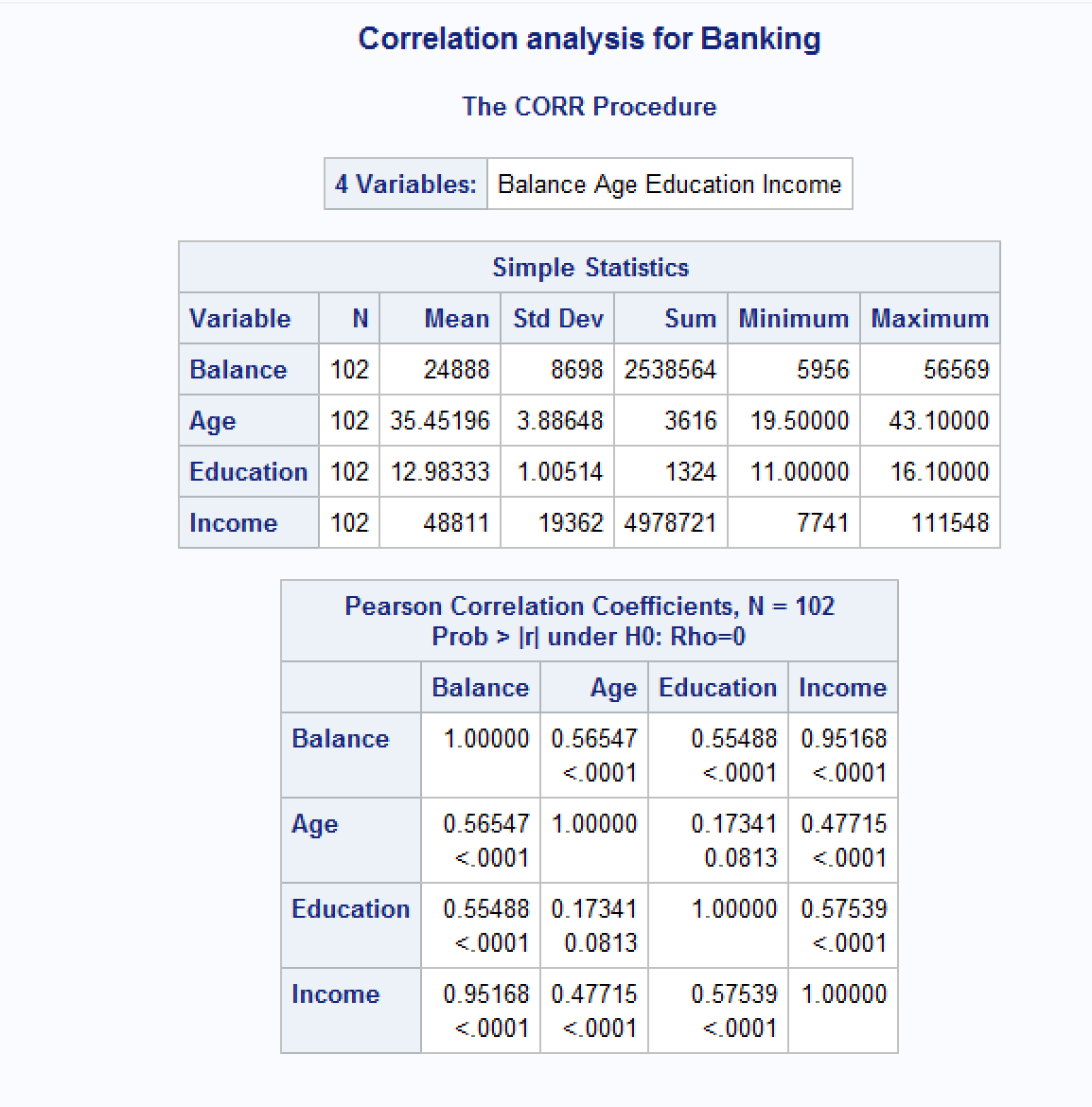
*TITLE 'Correlation analysis for Banking';*

*PROC CORR;*

*VAR Balance Age Education Income;*

*RUN;*

***Output:***

******

***Findings:***

*The Pearson Correlation Coefficients table tells us that there is a positive association between Balance and all three of the other variables with income having a very strong positive association with account balance at r=0.95168. Age(r=0.56547) and Education(r=0.55488) both have moderately strong positive associations with balance.*

1. What is the dependent variable and what are the independent variables in this regression analysis?

*Answer: The dependent variable in this regression analysis is Balance and the independent variables are Age, Education, and Income.*

1. Use SAS to fit a regression model to predict balance from age, education and income. Analyze the model parameters. Which predictors have a significant effect on balance? Use the t-tests on the parameters for alpha=0.05. Include the relevant regression output.

*Answer:*

***Code:***

*\*Regression Analysis: fits model to predict balance using age, education, and income;*

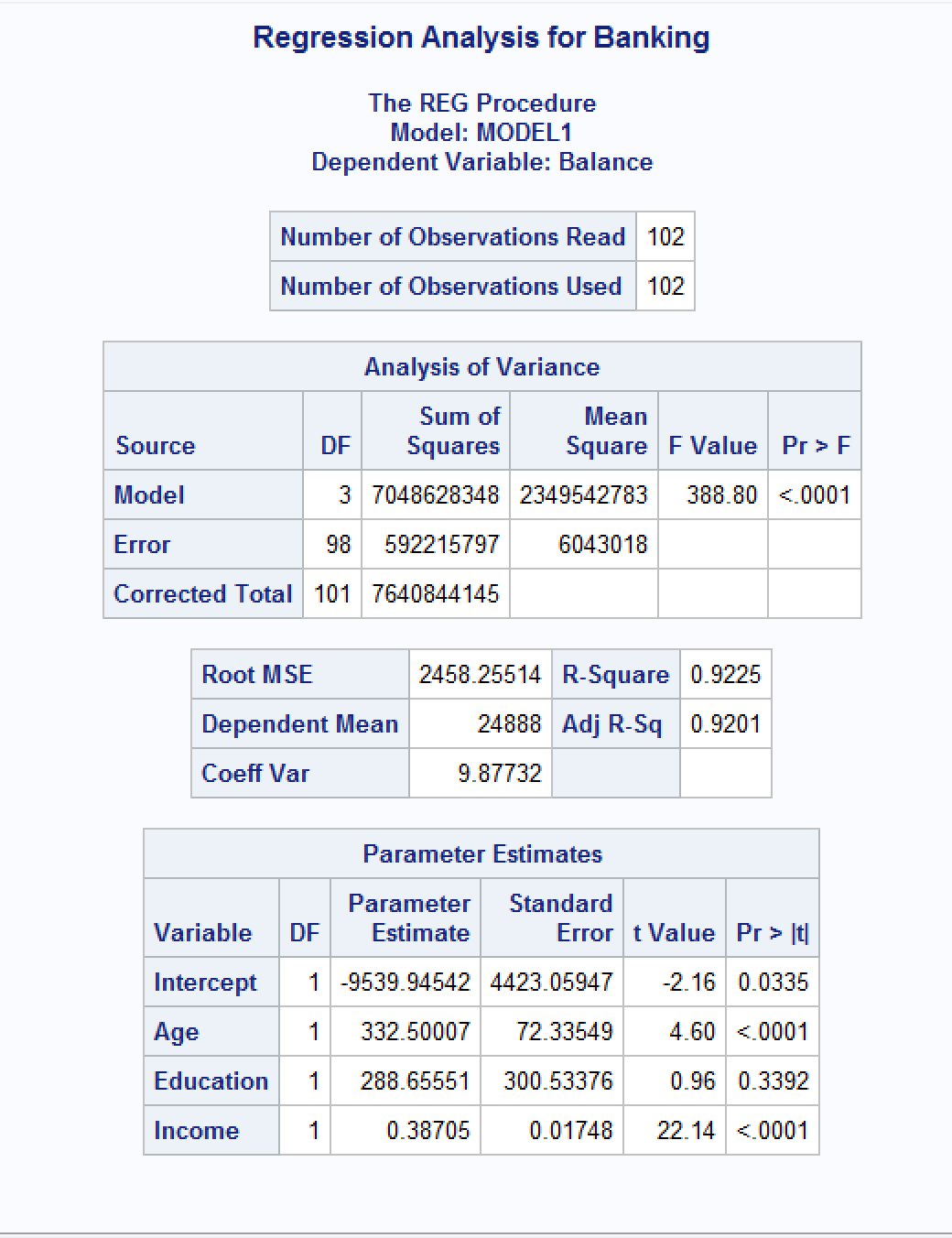
*TITLE 'Regression Analysis for Banking';*

*PROC REG data=banking;*

*model Balance=Age Education Income;*

*RUN;*

***Output:***

******

*The Parameter Estimates table in the above output which has a column of t-values for each of the predictor variable tells us that Age at alpha<.0001 is highly significant, as well as Income at alpha<.0001. Education has a t-value of alpha=0.3393 which is greater than 0.05 and therefore it is not significant as a predictor of account balance so we can remove it from our regression model.*

1. If one of the predictors is not significant, remove it from the model and refit the new regression model. Write the expression of the newly fitted regression model.

*Answer:*

***Code:***

*\*Regression Analysis with education removed;*

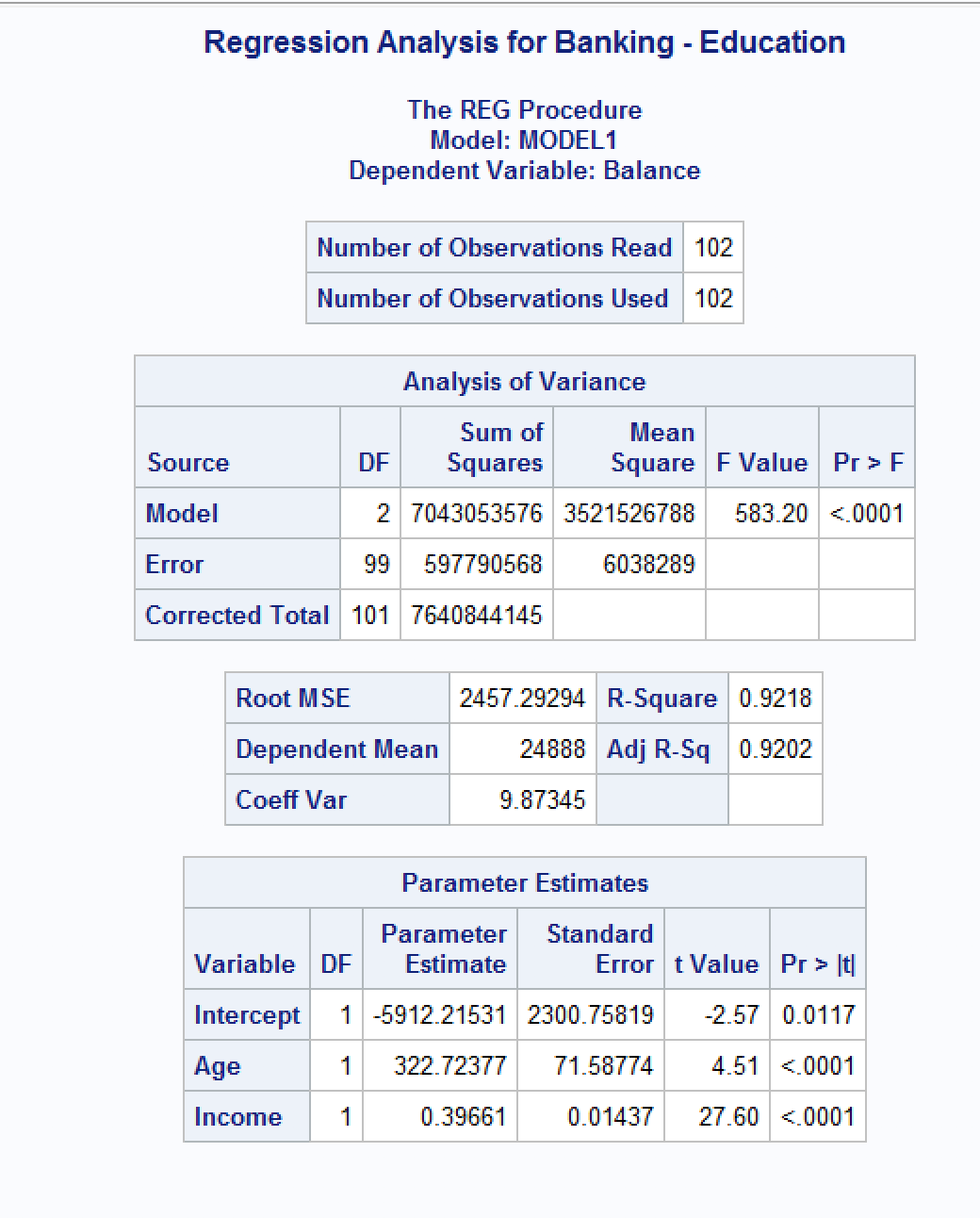
*TITLE 'Regression Analysis for Banking - Education';*

*PROC REG data=banking;*

*model Balance=Age Income;*

*RUN;*

***Output:***

******

*The final fitted model is: Balance = -5912.21531 + 322.72377(Age) + 0.39661(Income)*

1. Interpret the value of the parameters for the variables in the model.

*Answer:*

*I can interpret the value of the parameters for the variables in our final fitted model by calculating the standardized estimates to see which variables have the strongest effect on our y variable or Balance.*

***Code:***

*\*Regression Analysis with standardized estimates;*

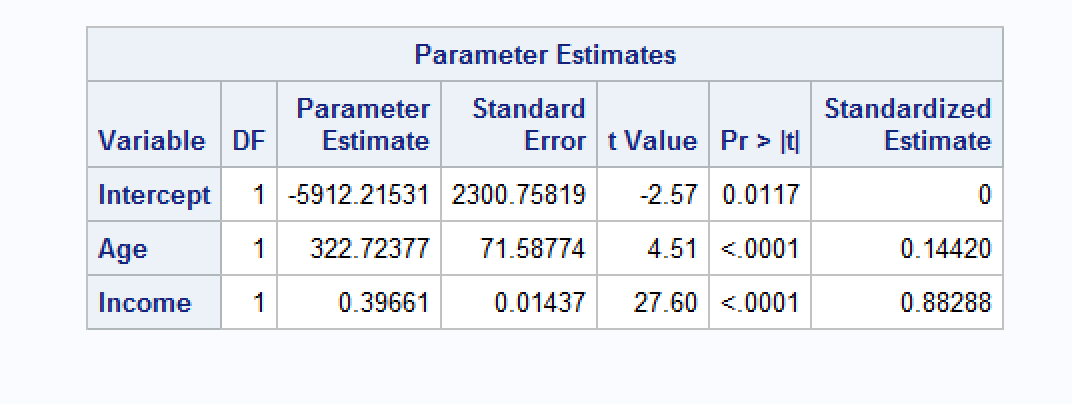
*TITLE 'Regression Analysis with Standardized Estimates';*

*PROC REG data=banking;*

*model Balance=Age Income /STB;*

*RUN;*

***Output:***

******

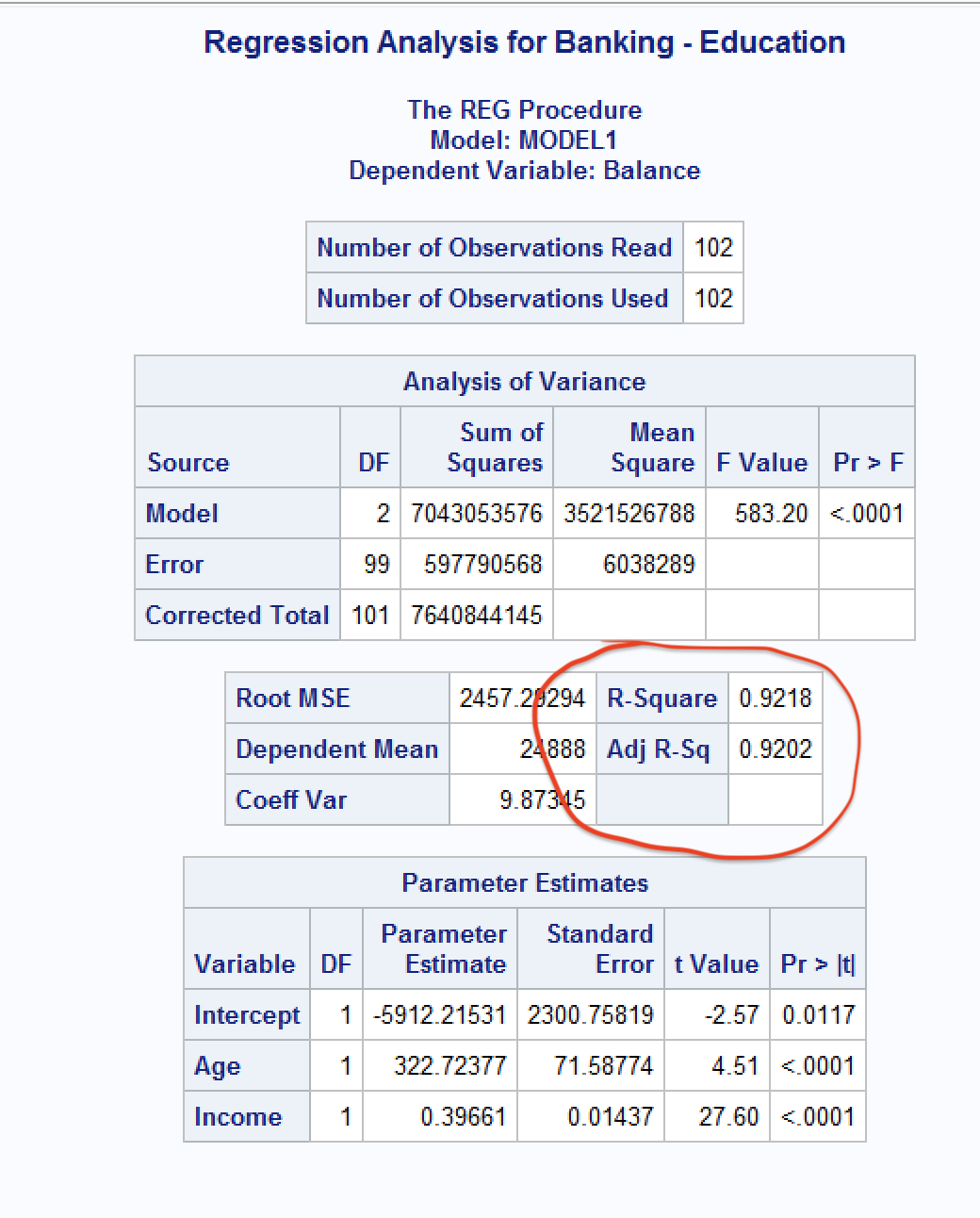
***Findings:***

*We can see from the right-most Standardized Estimate column in the above Parameter Estimates table that the median income(Income) variable is the most important variable or has the strongest effect on average bank balances(Balance) with a standardized estimate value of 0.88288. Age has a less strong effect with a value of 0.14420.*

1. Report the value for the R2and Adj-R2 coefficient and describe what it indicates. Include the portion of the output that includes the R2and Adj-R2 coefficient values.

*Answer:*

***Output:***

******

*The values for R2 = 0.9218 and* Adj-R2 = 0.9202. *These values both have similar meaning in that they both interpret the goodness of our model by telling us that approximately 92% of the variability in the average bank balances(Balance) can be explained using median age of the population(Age) and median income(Income). This is a good model because only approximately %8 of the variability of average bank balances cannot be explained by our model.*

1. According to census data, the population for a certain zip code area has median age equal to 34.8 years, median education equal to 12.5 years and median income equal to $42,401.

* Use the final model computed in step (f) above to compute the predicted average balance for the zip code area.

*Answer:*

*Using our final model: Balance = -5912.21531 + 322.72377(Age) + 0.39661(Income), we can plug in the values above to compute the average balance:*

*Balance = -5912.21531 + 322.72377(34.8) + 0.39661(42401)*

*Balance = -5912.21531 + 11,230.7872 + 16,816.6606*

*Predicted Average Balance for the zip code area = $22,135.2325*

* If the observed average balance for the zip code area is $21,572, what’s the model prediction error?

21572 = *22,135.2325 + e*

*Prediction error = 21572 - 22,135.2325*

*Prediction error = $-563.2325*

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

*Answer: See answers and output above*