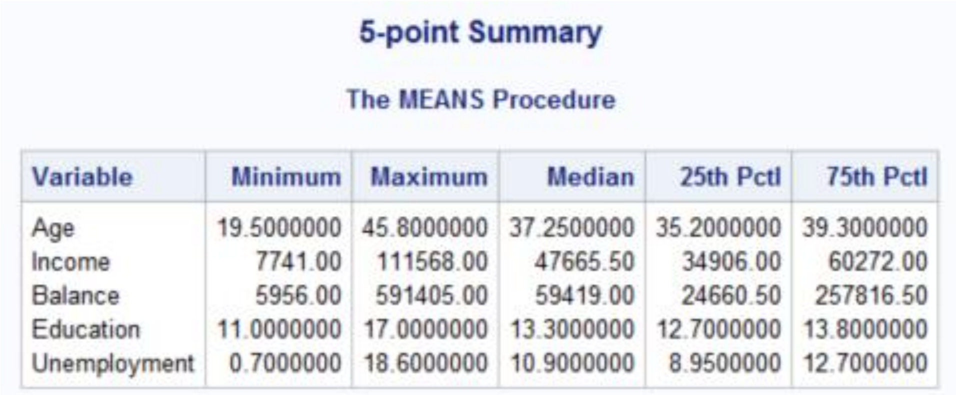
Ian Mulchrone

DSC 423

Homework 2

Problem 1

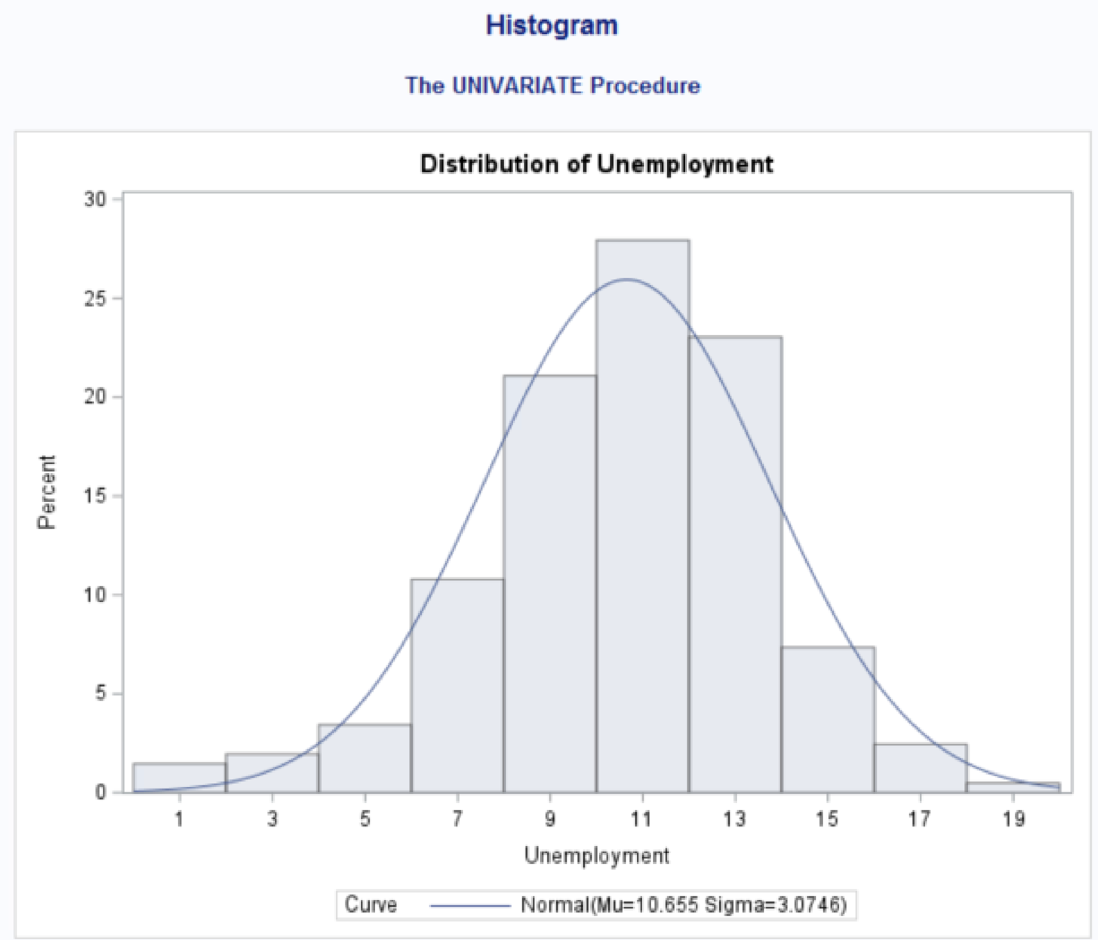
a)



5-point Summary

The MEANS Procedure

Variable	Minimum	Maximum	Median	25th Pctl	75th Pctl
Age	19.5000000	45.8000000	37.2500000	35.2000000	39.3000000
Income	7741.00	111568.00	47665.50	34906.00	60272.00
Balance	5956.00	591405.00	59419.00	24660.50	257816.50
Education	11.0000000	17.0000000	13.3000000	12.7000000	13.8000000
Unemployment	0.7000000	18.6000000	10.9000000	8.9500000	12.7000000

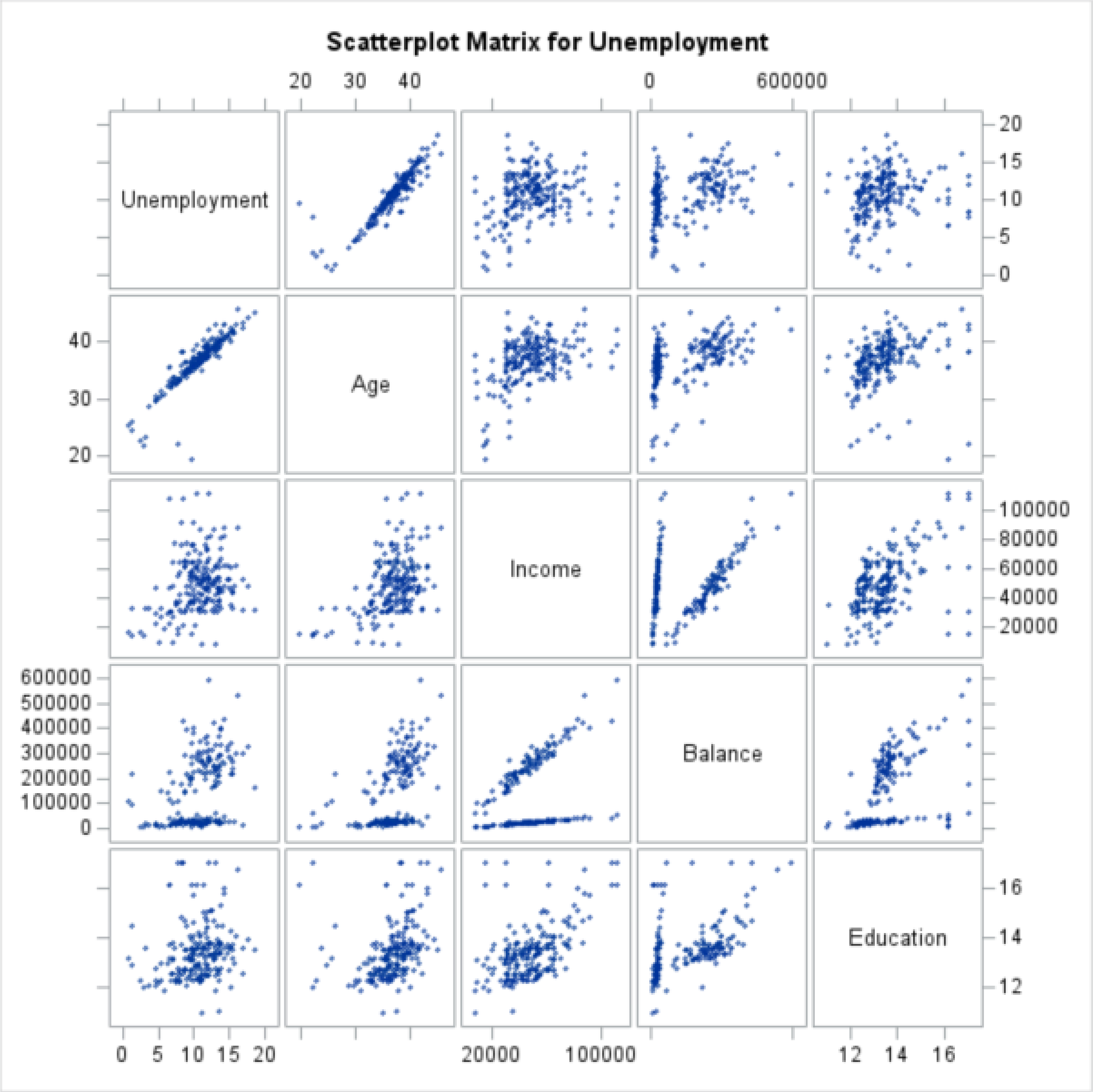


Histogram

The UNIVARIATE Procedure

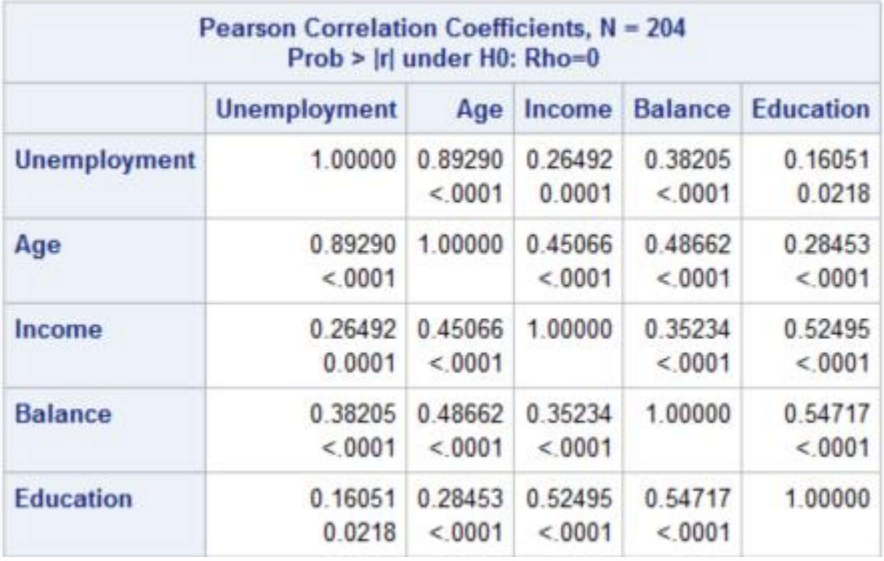
The distribution of unemployment appears to be normal with a mean of 10.655% unemployment and most zip codes falling between 8-14% unemployment. This is reinforced when looking at the five number summary, with a median unemployment rate of 10.9% which is close to the mean, indicating normality, and an IQR of 8.95%-12.7% supporting the results of the histogram. There appears to be outliers at the left tail because with a standard deviation of 3.075, any zip code with unemployment below about 1.4% would be an outlier further than three standard deviations from the mean. We know there is at least one at 0.7%, but we will need to do additional analysis to see if there are others. There are no outliers at the right tail, since the maximum observed unemployment is 18.6% which falls within three standard deviations.

b)



Age appears to have a high correlation with unemployment, with the scatterplot showing a linear relationship between the two. Income and education both have a positive, medium level correlation with unemployment. Balance is more difficult to say, since there is a very high correlation when Balance is close to 0, but shows a medium correlation with any non-zero balance.

c)



Pearson Correlation Coefficients, N = 204
Prob > |r| under H0: Rho=0

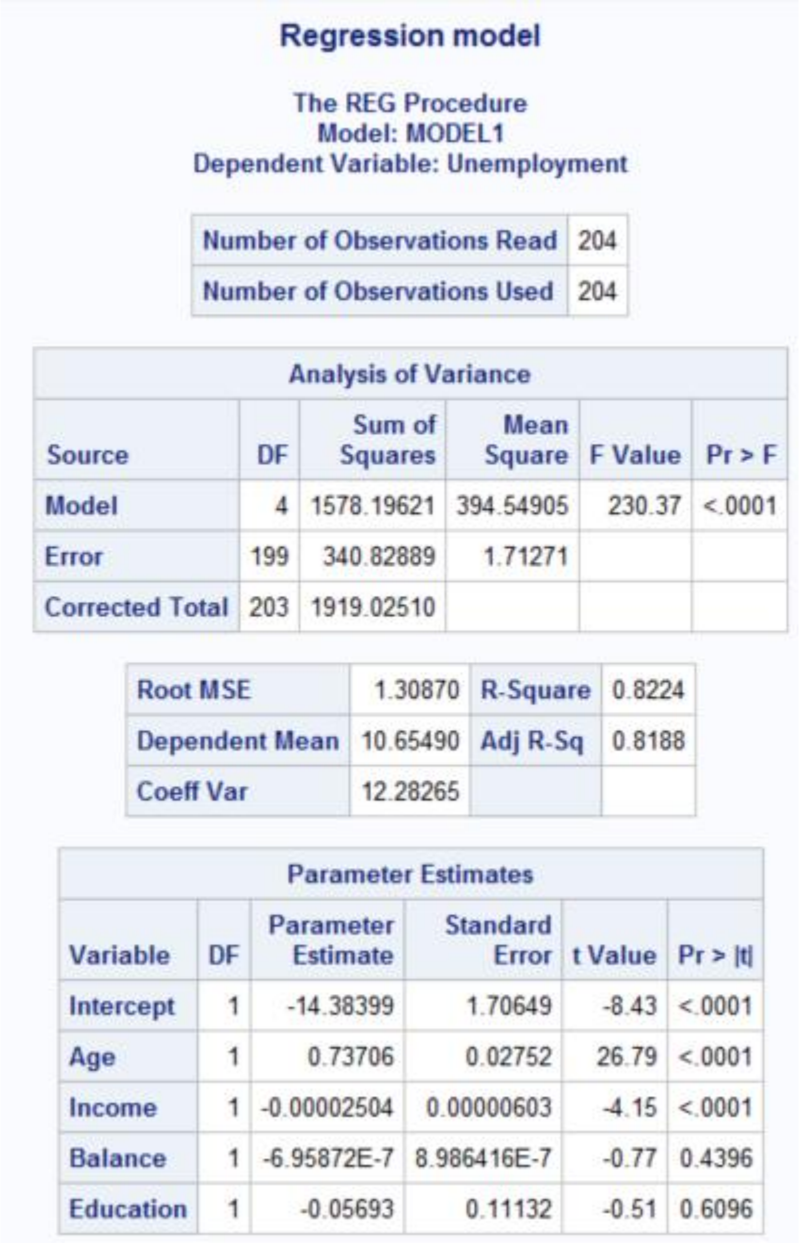
	Unemployment	Age	Income	Balance	Education
Unemployment	1.00000	0.89290 <.0001	0.26492 0.0001	0.38205 <.0001	0.16051 0.0218
Age	0.89290 <.0001	1.00000	0.45066 <.0001	0.48662 <.0001	0.28453 <.0001
Income	0.26492 0.0001	0.45066 <.0001	1.00000	0.35234 <.0001	0.52495 <.0001
Balance	0.38205 <.0001	0.48662 <.0001	0.35234 <.0001	1.00000	0.54717 <.0001
Education	0.16051 0.0218	0.28453 <.0001	0.52495 <.0001	0.54717 <.0001	1.00000

As shown in the scatterplot matrix, Age has a high positive correlation with Unemployment. It appears I overestimated the strength of Income and Education, with both having very low correlation values below 0.3. Balance also has a low correlation value of 0.38. Looking back on the scatterplot now this value makes sense, since the zip codes with an average bank balance close to 0 covers a wide range of Unemployment outcomes (about 2-17%).

d)

The dependent variable is Unemployment and the independent variables are Age, Income, Balance, and Education.

e)



Regression model

The REG Procedure
Model: MODEL1
Dependent Variable: Unemployment

Number of Observations Read	204
Number of Observations Used	204

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1578.19621	394.54905	230.37	<.0001
Error	199	340.82889	1.71271		
Corrected Total	203	1919.02510			

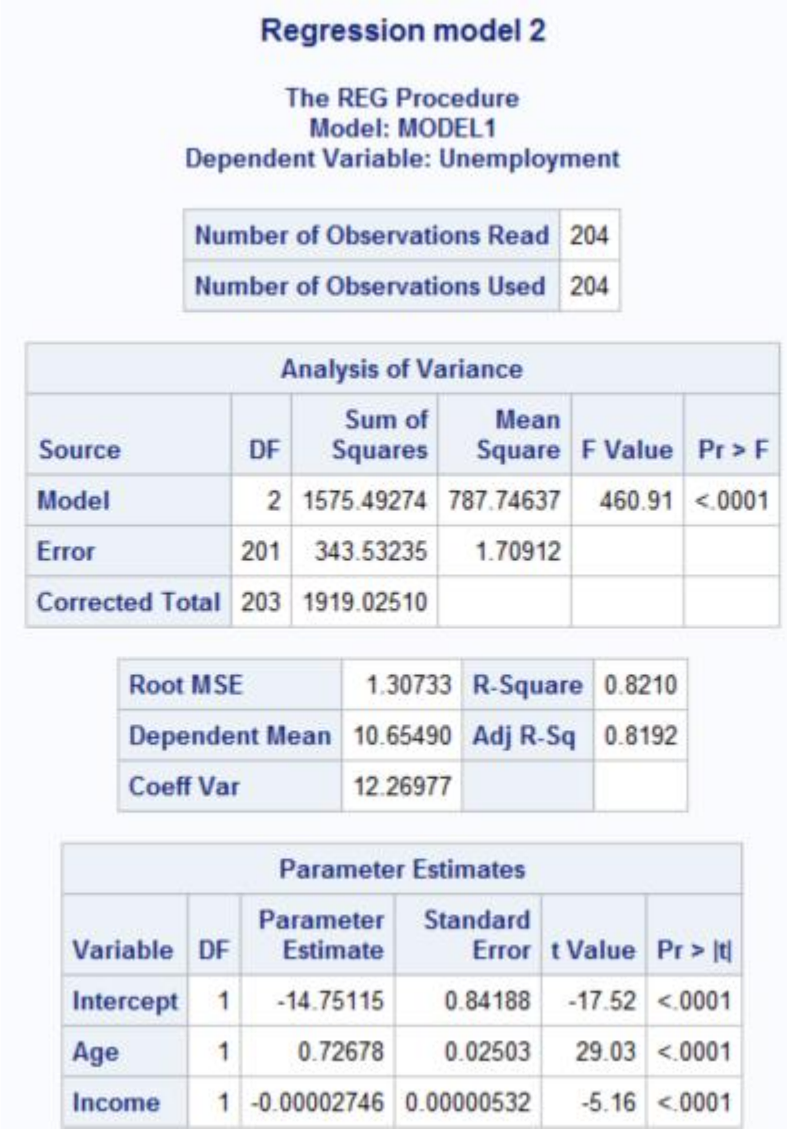
Root MSE	1.30870	R-Square	0.8224
Dependent Mean	10.65490	Adj R-Sq	0.8188
Coeff Var	12.28265		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-14.38399	1.70649	-8.43	<.0001
Age	1	0.73706	0.02752	26.79	<.0001
Income	1	-0.00002504	0.00000603	-4.15	<.0001
Balance	1	-6.95872E-7	8.986416E-7	-0.77	0.4396
Education	1	-0.05693	0.11132	-0.51	0.6096

Age and income have a significant effect on unemployment since the both have a P-value < 0.0001, which falls under the 95% significance threshold 0.05. Balance and education both have P-values well above 0.05, meaning they are not statistically significant in the model.

f)



Regression model 2

The REG Procedure
Model: MODEL1
Dependent Variable: Unemployment

Number of Observations Read	204
Number of Observations Used	204

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1575.49274	787.74637	460.91	<.0001
Error	201	343.53235	1.70912		
Corrected Total	203	1919.02510			

Root MSE	1.30733	R-Square	0.8210
Dependent Mean	10.65490	Adj R-Sq	0.8192
Coeff Var	12.26977		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-14.75115	0.84188	-17.52	<.0001
Age	1	0.72678	0.02503	29.03	<.0001
Income	1	-0.00002746	0.00000532	-5.16	<.0001

Unemployment = -14.75115 + 0.72678*Age – 0.00002746*Income

g)

For every 1-year increase in median age, unemployment will increase by 0.72678%.

For every 1$ increase in median income, unemployment will decrease by 0.00002746%.

h)

The R-Square value 0.821 means that 82.1% of the variation in unemployment can be explained by median age and median income.

The Adj R-sq value .8192 means that 81.92% of the variation in unemployment can be explained by median age and median income.

i)

i)

Unemployment = -14.75115 + 0.72678*44.2 – 0.00002746*51324

Unemployment = 15.963%

In a zip code with a median age of 44.2 years, median education of 11.5 years, median income of $51,324, and average bank balance of $34,200, we can expect the unemployment rate to be 15.963%

ii)

If the observed unemployment for the zip code is 13.5%, the model prediction error is 2.463%.

j)

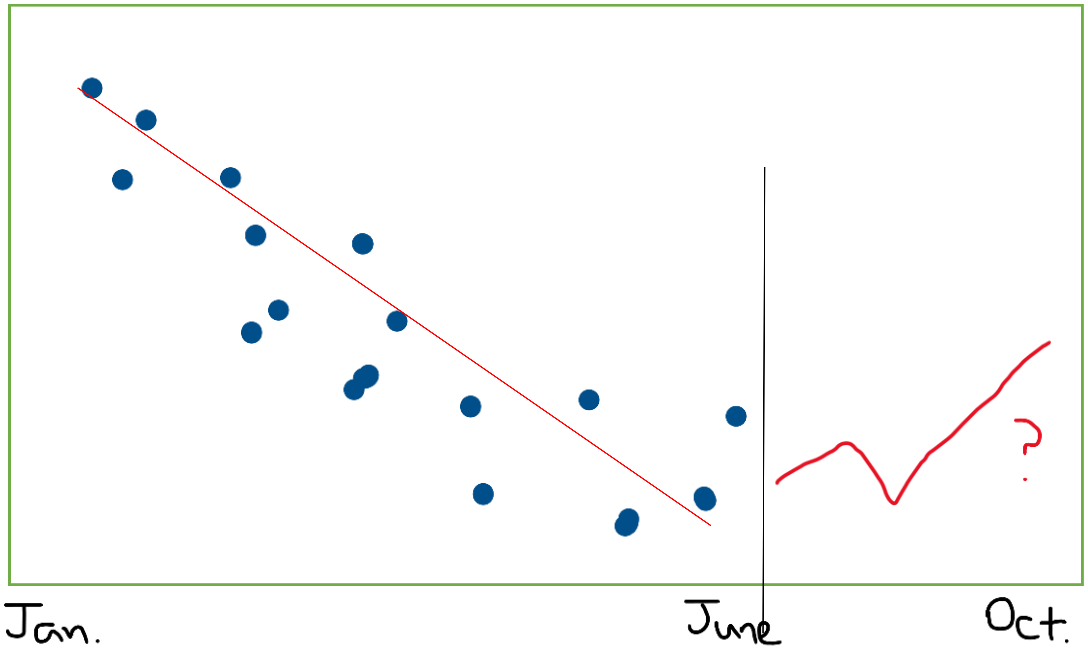
```
*Import the dataset;
PROC IMPORT datafile="unemployment.txt" out=unemployment replace;
delimiter='09'x;
getnames=YES;
datarow=2;
RUN;

*prints the dataset;
TITLE "Dataset - Unemployment";
PROC PRINT;
RUN;

*5-point summary ;
TITLE "5-point Summary";
```

```
PROC MEANS min max median p25 p75;
VAR Age Income Balance Education Unemployment;
RUN;

*Histogram;
TITLE "Histogram";
PROC UNIVARIATE normal;
VAR Unemployment;
histogram / normal (mu = est sigma = est);
RUN;

*Scatterplots;
TITLE "Scatterplots";
PROC GPLOT;
PLOT Unemployment*(Age Income Balance Education);
RUN;

*Scatterplot Matrix;
TITLE "Scatterplot Matrix for Unemployment";
PROC SGSCATTER;
MATRIX Unemployment Age Income Balance Education;
RUN;

*Correlation values;
TITLE "Correlation values";
PROC CORR;
VAR Unemployment Age Income Balance Education;
RUN;
```

```
*Regression model;

TITLE "Regression model";

PROC REG;

MODEL Unemployment=Age Income Balance Education;

RUN;


*Regression model 2;

TITLE "Regression model 2";

PROC REG;

MODEL Unemployment=Age Income;

RUN;
```

Problem 2

1. The prediction error was higher for October compared to May because October was outside of the dataset, so we were using extrapolation to predict that price point. Because May was within the dataset, thus using interpolation, the prediction was more accurate.



Using the above drawing, even if we see a negative trend in the data, whatever comes after the last observation in the dataset is still unknown. It is possible that there was a shift after the data ends in June, but we do not know so all we can do is guess when extrapolating data.

2

a. K = 2
b. K = 3

3

The three errors we discussed were Sum of Squares Error (SSE), Mean Square Error (MSE), and Root Mean Square Error (MRSE).

Problem 3

1.

I believe figure 2 will produce a more accurate prediction because it has a linear shape with no discernable outliers. Figure 1 does not show a linear relationship and has a large gap in the data with many observations at 0 on the x-axis, which will increase the error when making predictions.

2.

Figure 1 has a very low positive correlation. The line of best fit shows a positive relationship, but the I believe the correlation coefficient would be near 0.

Figure 2 has a medium negative correlation. The observations deviate too far from the line of best fit to be a strong correlation, but it clearly shows a negative trend.