SAS Certificate Portfolio

Eric Fernandez

Spring 2020

Friday, April 17, 2020

STA 3024

Project Phase A

Eric Fernandez

Motivation

Winemaking is a lengthy process that involves several factors such as environmental conditions, chemical properties of materials used, type of grape and others. It takes years of expertise to know how to produce and determine what makes a good wine and, even after a wine is produced, the flavor is constantly changing over time. Quality of wines is usually determined by sommeliers. Based on the information provided by the winemaker and the reviews by sommeliers, is there a way to meet the requirements of what makes a good wine and, by doing so, produce better wines of different varieties? Over the semester, I hope to find answers to some of the questions below:

1)Is there a way to determine what makes a great wine based on specific descriptors?

2)If so, is there any relationship between good quality and price? Can you predict how much a wine will cost based on a review by a sommelier?

3)Can we identify regions that produce better wine than others?

4)If there are regions that produce better wine than others, are there special conditions in these regions that help produce these results?

5)If there are special conditions in these regions, are these conditions replicable in regions that do not produce wine that could potentially produce similar quality wine?

By addressing these questions, I look to understand what are the main factors that determine the quality in wine.

Data Description

The dataset I am using was collected by Zack Thoutt scraping the website WineEnthusiast during the week of June 15th, 2017. The code used to scrap the data can be found here.

The columns describe the following attributes for every data point:

- *Points*: the number of points WineEnthusiast rated the wine on a scale of 1-100. WineEnthusiasts only post reviews for wines that score >=80.

- *Variety*: the type of grapes used to make the wine (ie Pinot Noir)
- *Description*: a few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.
- *Country*: the country that the wine is from
- *Province*: the province or state that the wine is from
- *Region 1*: the wine growing area in a province or state (ie Napa)
- *Region 2*: sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank
- *Winery*: the winery that made the wine
- *Designation*: the vineyard within the winery where the grapes that made the wine are from
- *Price*: the cost for a bottle of the wine

These descriptions were created by the Zack Thoutt. The dataset was downloaded as a csv file from Kaggle(https://www.kaggle.com/zynicide/wine-reviews). By having the price, location and description from sommeliers, I will be able to answer questions 1 through 3. 4 will require more digging into the processes and conditions.

SAS Implementation

Some of the issues I encountered in this dataset were missing values for region fields and incorrect formatting when importing. Incorrect formatting was due to the fact that several lines were split into two or more in the original dataset file which caused SAS to have errors when reading the csv file. By concatenating the lines that were separated, this error was solved. Region data is not relevant to questions 1 and 2 but questions 3 and 4 need region fields so wines with no region data will not be considered for these particular questions.

Code:

```
/* Eric Fernandez Project-Phase A*/
/* I certify that the SAS code given is my original and exclusive work*/

/* To read the file:
```

Create a new folder.

Upload 'winemag-data_first150k.csv' to the folder

Right click on 'winemag-data_first150k.csv' and select Properties

Copy the path name and paste to the filename statement below

Add a slash and the file name to the end of the path

```
*/
FILENAME CSV "~/datasets/winemag-data_first150k.csv" TERMSTR=LF;
/** Import the CSV file.  **/
PROC IMPORT DATAFILE=CSV
            OUT=WineReviews
            DBMS=CSV
            REPLACE;
RUN;
/*Print out the first 20 reviews out of 150,000 reviews*/
proc print data=WineReviews(obs=20);
run;
```

Output:

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
1	0	US	This tremendous 100% varietal wine hails from Oakville and was aged over three years in oak. Juicy red-cherry fruit and a compelling hint of caramel greet the	Martha's Vineyard	96	235	California	Napa Valley	Napa	Cabernet Sauvignon	Heitz

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			palate, framed by elegant, fine tannins and a subtle minty tone in the background. Balanced and rewarding from start to finish, it has years ahead of it to develop further nuance. Enjoy 2022–2030.								
2	1	Spain	Ripe aromas of fig, blackberry and cassis are softened and sweetened by a slathering of oaky chocolate and vanilla. This is full, layered, intense and cushioned on the palate, with rich	Carodorum Selección Especial Reserva	96	110	Northern Spain	Toro		Tinta de Toro	Bodega Carmen Rodríguez

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			flavors of chocolaty black fruits and baking spices. A toasty, everlasting finish is heady but ideally balanced. Drink through 2023.								
3	2	US	Mac Watson honors the memory of a wine once made by his mother in this tremendously delicious, balanced and complex botrytised white. Dark gold in color, it layers toasted hazelnut, pear compote and orange peel flavors, reveling in the succulence of its 122 g/L of	Special Selected Late Harvest	96	90	California	Knights Valley	Sonoma	Sauvignon Blanc	Macauley

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			residual sugar.								
4	3	US	This spent 20 months in 30% new French oak, and incorporates fruit from Ponzi's Aurora, Abetina and Madrona vineyards, among others. Aromatic, dense and toasty, it deftly blends aromas and flavors of toast, cigar box, blackberry, black cherry, coffee and graphite. Tannins are polished to a fine sheen, and frame a finish loaded with dark chocolate and espresso. Drink now	Reserve	96	65	Oregon	Willamette Valley	Willamette Valley	Pinot Noir	Ponzi

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			through 2032.								
5	4	France	This is the top wine from La Bégude, named after the highest point in the vineyard at 1200 feet. It has structure, density and considerable acidity that is still calming down. With 18 months in wood, the wine has developing an extra richness and concentration. Produced by the Tari family, formerly of Château Giscours in Margaux, it is a wine made for aging. Drink from 2020.	La Brûlade	95	66	Provence	Bandol		Provence red blend	Domaine de la Bégude

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
6	5	Spain	Deep, dense and pure from the opening bell, this Toro is a winner. Aromas of dark ripe black fruits are cool and moderately oaked. This feels massive on the palate but sensationally balanced. Flavors of blackberry, coffee, mocha and toasty oak finish spicy, smooth and heady. Drink this exemplary Toro through 2023.	Numanthia	95	73	Northern Spain	Toro		Tinta de Toro	Numanthia
7	6	Spain	Slightly gritty black-fruit aromas include a sweet note of pastry along with a hint of prune. Wall-to-	San Román	95	65	Northern Spain	Toro		Tinta de Toro	Maurodos

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			wall saturation ensures that all corners of one's mouth are covered. Flavors of blackberry, mocha and chocolate are highly impressive and expressive, while this settles nicely on a long finish. Drink now through 2024.								
8	7	Spain	Lush cedary black-fruit aromas are luxe and offer notes of marzipan and vanilla. This bruiser is massive and tannic on the palate, but still lush and friendly. Chocolate is a key flavor, while baked	Carodorum Único Crianza	95	110	Northern Spain	Toro		Tinta de Toro	Bodega Carmen Rodríguez

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			berry and cassis flavors are hardly wallflowers. On the finish, this is tannic and deep as a sea trench. Drink this saturated black-colored Toro through 2023.								
9	8	US	This re-named vineyard was formerly bottled as deLancellotti. You'll find striking minerality underscoring chunky black fruits. Accents of citrus and graphite comingle, with exceptional midpalate concentration. This is a wine to cellar, though it is already quite	Silice	95	65	Oregon	Chehalem Mountains	Willamette Valley	Pinot Noir	Bergström

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			enjoyable. Drink now through 2030.								
10	9	US	The producer sources from two blocks of the vineyard for this wine—one at a high elevation, which contributes bright acidity. Crunchy cranberry, pomegranate and orange peel flavors surround silky, succulent layers of texture that present as fleshy fruit. That delicately lush flavor has considerable length.	Gap's Crown Vineyard	95	60	California	Sonoma Coast	Sonoma	Pinot Noir	Blue Farm
11	10	Italy	Elegance, complexity and structure come together in	Ronco della Chiesa	95	80	Northeastern Italy	Collio		Friulano	Borgo del Tiglio

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			this drop-dead gorgeous winethat ranks among Italy's greatest whites. It opens with sublime yellow spring flower, aromatic herb and orchard fruit scents. The creamy, delicious palate seamlessly combines juicy white peach, ripe pear and citrus flavors while white almond and savory mineral notes grace the lingering finish.								
12	11	US	From 18-year-old vines, this supple well-balanced effort blends flavors of	Estate Vineyard Wadensvil Block	95	48	Oregon	Ribbon Ridge	Willamette Valley	Pinot Noir	Patricia Green Cellars

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			mocha, cherry, vanilla and breakfast tea. Superbly integrated and delicious even at this early stage, this wine seems destined for a long and savory cellar life. Drink now through 2028.								
13	12	US	A standout even in this terrific lineup of 2015 releases from Patricia Green, the Weber opens with a burst of cola and tobacco scents and accents. It continues, subtle and detailed, with flavors of oranges, vanilla, tea and milk chocolate discreetly	Weber Vineyard	95	48	Oregon	Dundee Hills	Willamette Valley	Pinot Noir	Patricia Green Cellars

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			threaded through ripe blackberry fruit.								
14	13	France	This wine is in peak condition. The tannins and the secondary flavors dominate this ripe leather-textured wine. The fruit is all there as well: dried berries and hints of black-plum skins. It is a major wine right at the point of drinking with both the mature flavors and the fruit in the right balance.	Château Montus Prestige	95	90	Southwest France	Madiran		Tannat	Vignobles Brumont
15	14	US	With its sophisticated mix of mineral, acid and tart fruits, this seductive effort pleases	Grace Vineyard	95	185	Oregon	Dundee Hills	Willamette Valley	Pinot Noir	Domaine Serene

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			from start to finish. Supple and dense, it's got strawberry, blueberry, plum and black cherry, a touch of chocolate, and that underlying streak of mineral. All these elements are in good proportion and finish with an appealing silky texture. It's delicious already, but give it another decade for full enjoyment. Drink now through 2028.								
16	15	US	First made in 2006, this succulent luscious Chardonnay is all about minerality.	Sigrid	95	90	Oregon	Willamette Valley	Willamette Valley	Chardonnay	Bergström

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			It's got a rich core of butterscotch and the seemingly endless layers of subtle flavors that biodynamic farming can bring. It spends 18 months on the lees prior to bottling. Drink now through 2028.								
17	16	US	This blockbuster, powerhouse of a wine suggests blueberry pie and chocolate as it opens in the glass. On the palate, it's smooth and seductively silky, offering complex cedar, peppercorn and peppery oak seasonings	Rainin Vineyard	95	325	California	Diamond Mountain District	Napa	Cabernet Sauvignon	Hall

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			amidst its dense richness. It finishes with finesse and spice.								
18	17	Spain	Nicely oaked blackberry, licorice, vanilla and charred aromas are smooth and sultry. This is an outstanding wine from an excellent year. Forward barrel-spice and mocha flavors adorn core blackberry and raspberry fruit, while this runs long and tastes vaguely chocolaty on the velvety finish. Enjoy this top-notch Tempranillo through 2030.	6 Años Reserva Premium	95	80	Northern Spain	Ribera del Duero		Tempranillo	Valduero

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
19	18	France	Coming from a seven-acre vineyard named after the dovecote on the property, this is a magnificent wine. Powered by both fruit tannins and the 28 months of new wood aging, it is darkly rich and with great concentration. As a sign of its pedigree, there is also elegance here, a restraint which is new to this wine. That makes it a wine for long-term aging. Drink from 2022.	Le Pigeonnier	95	290	Southwest France	Cahors		Malbec	Château Lagrézette
20	19	US	This fresh and lively medium-bodied wine is beautifully	Gap's Crown Vineyard	95	75	California	Sonoma Coast	Sonoma	Pinot Noir	Gary Farrell

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			crafted, with cherry blossom aromas and tangy acidity. Layered and seductive, it offers a crisp mix of orange peel, cherry, pomegranate and baking spice flavors that are ready for the table or the cellar.								

Future Direction

For the next steps, in order to analyze and learn about the dataset more I planned to:

- Find models to predict the variables I am looking for and compared them for this use case.
- Find better ways to visualize the data.
- Research about processes used to create wine and conditions of regions with good wine quality in this dataset.
- Create a dictionary of most common words used by sommeliers.
- Find regions that have similar conditions to the ones used in the dataset that currently do not produce wine to check if there are new regions that could potentially produce similar kind of wines.

We, the project team members, certify that the percentage of the effort listed by each of our names below is an accurate account of the original effort contributed by each team member in the producing of this project and report:

Name (Printed)	Percent of Total Effort	Statistics Major?
Eric Fernandez	100	No

Project Phase B

Introduction

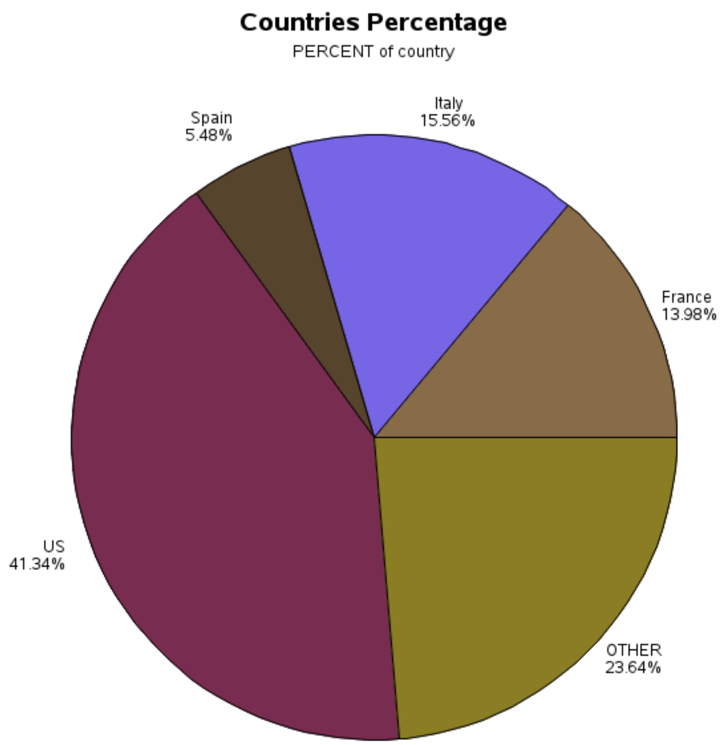
My main motivation for this project is to find if there are significant relationships between countries and quality of wine, varieties and price, and quality and price. This analysis can help produce better wines, predict numerically how good it would be and decide which factors are most important when producing a high quality wine.
The variables of interest used for this phase are:

- *Points*: the number of points WineEnthusiast rated the wine on a scale of 1-100. WineEnthusiasts only post reviews for wines that score >=80.
- *Country*: the country that the wine is from
- *Price*: the cost for a bottle of the wine
- *Variety*: the type of grapes used to make the wine (ie Pinot Noir)

In this phase, I will explore the wine review dataset by using graphical displays and numerical summaries to find if there is a relationship between price and quality and how each country category compares to each other.

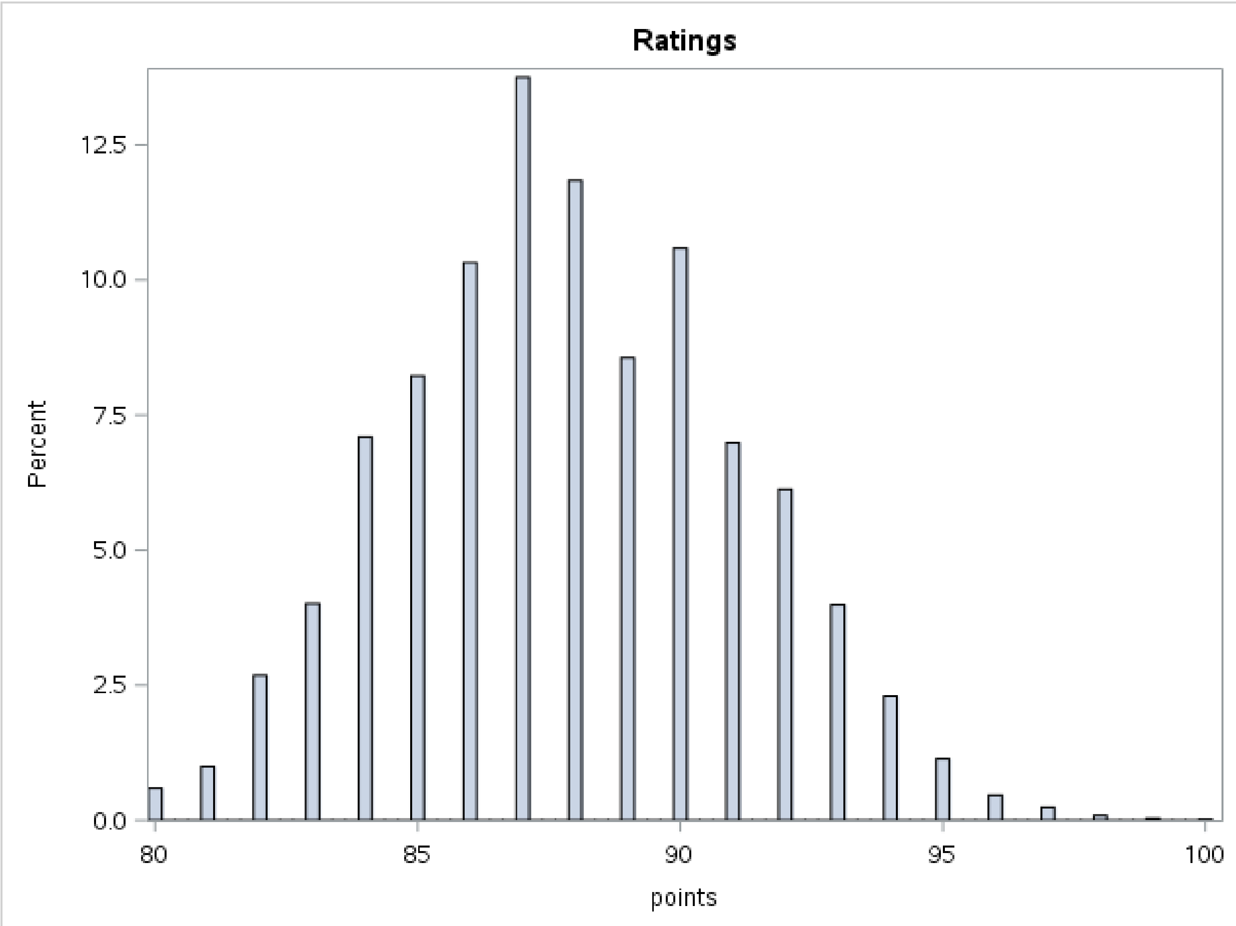
Data Exploration via Graphical Display

a)



The majority of the wines reviewed are from the US(41.34%). There are similar amounts of wines reviews of wines coming from France and Italy. The "Other Countries" category consists of a of 47 countries. This group includes: Albania, Argentina, Austria, Australia, Bosnia, Brazil, Bulgaria, Canada, Chile, China, Croatia, Cyprus, Czech Republic, Egypt, England, Georgia, Germany, Greece, Hungary, India, Israel, Japan, Lebanon, Lithuania, Luxembourg, Macedonia, Moldovia, Montenegro, Morocco, New Zealand, Portugal, Romania, Serbia, Slovakia, Slovenia, South Africa, Switzerland, Tunisia, Turkey, Ukraine and Uruguay.

b)

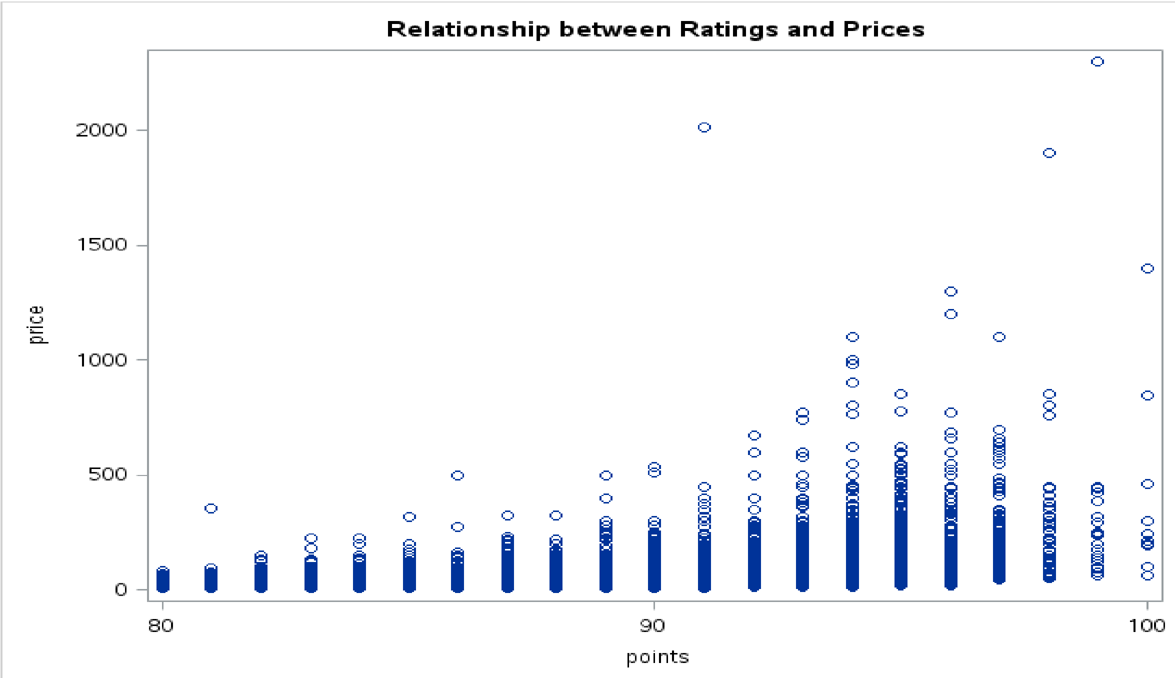


The ratings follow slightly a bell shaped right-skewed distribution. The center is at 87. Most of the wine ratings are above the median.

d)

This dataset contains outliers in every country category. Spain, U.S. and "Other Countries" have outliers above the third quartile while countries like Italy and France have outliers below and above the first and third quartile respectively. The highest median is from France, around 89 points while the "Other Countries" category and Spain have the lowest medians.

e)



The density of points displayed in the graph suggests that there is a high amount of wines priced below $500 dollars. Wines with ratings above 90 tend to be more expensive with one outlier reaching above $2,000 dollars.

Data Exploration via Numerical Summaries

a)

Frequency of Country

The FREQ Procedure

country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
France	21098	13.98	21098	13.98
Italy	23478	15.56	44576	29.53
Other	35689	23.65	80265	53.18
Spain	8268	5.48	88533	58.66
US	62397	41.34	150930	100.00

This table summarizes numerically the graphical representation of the countries in the pie chart of in the data exploration part showing once again that majority of wines reviewed in this dataset come from the U.S.

b)

Descriptive Analysis of Prices of Wine

The MEANS Procedure

Analysis Variable : price				
N	Mean	Std Dev	Minimum	Maximum
137235	33.1314825	36.3225362	4.0000000	2300.00

Wines in this dataset can reach the price of $2,300.00 dollars and can be as low as $4.00 dollars. The mean of the dataset being $33.13 dollars. The standard deviation is large because the min and max are far apart due to outliers.

d)A descriptive analysis of prices of wine per variety is included in Table 1 of the appendix. This analysis shows that the mean, minimum, maximum value vary largely between wine varieties. Not all wine varieties are included in Table 1 however, it shows a representation of the sporadic changes in price.

e)

Relationship between Rating and Price

The CORR Procedure

2 Variables:	points price

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
points	150930	87.88842	3.22239	13264999	80.00000	100.00000
price	137235	33.13148	36.32254	4546799	4.00000	2300

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	points	price
points	1.00000 150930	0.45986 <.0001 137235
price	0.45986 <.0001	1.00000

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	points	price
	137235	137235

The correlation coefficient obtained is 0.45986 suggesting that there is a moderately positive correlation between price and quality points.

Conclusion

The graphical analysis suggests that the majority of the wines reviewed in this dataset are from the U.S.(41.34 percent) with the "Other Countries" category having the second highest percentage(23.64%), followed by France(13.98 percent) and Italy(15.56 percent) with similar percentages and finally Spain(5.48) with the lowest percentage. We can also see that the median for the quality of wines in this dataset ranging from 80 to 100 is 87. The results suggest that there is a moderately positive correlation between price and quality with some outliers surpassing $2,000. For the next phase of the project, I will run an ANOVA test on the quality points variable in order to observe if there exists a statistical significance in the difference between means that can determine if there is a country that produces better wine on average. From there, I would like to analyze the quality of wine based on the regions of the best country/countries and check whether there are special conditions these wines are prepared.

We, the project team members, certify that the percentage of the effort listed by each of our names below is an accurate account of the original effort contributed by each team member in the producing of this project and report:

Name (Printed)	Percent of Total Effort	Statistics Major?
Eric Fernandez	100	No

Appendix

Table1

Descriptive analysis of Prices of Wine per Variety

The MEANS Procedure

Analysis Variable : price						
variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Agiorgitiko	120	117	19.2991453	10.0243367	8.0000000	65.0000000
Aglianico	317	259	33.1698842	19.0465083	6.0000000	130.0000000
Aidani	1	1	27.0000000	.	27.0000000	27.0000000
Airen	6	6	8.8333333	0.7527727	8.0000000	10.0000000
Albana	17	15	33.9333333	19.2816221	8.0000000	66.0000000
Albariño	537	530	19.9924528	7.6472279	10.0000000	110.0000000
Albarossa	1	1	40.0000000	.	40.0000000	40.0000000
Albarín	1	1	15.0000000	.	15.0000000	15.0000000
Aleatico	11	10	37.9000000	7.4304180	30.0000000	50.0000000
Alfrocheiro	18	18	24.0000000	11.9114379	11.0000000	40.0000000
Alicante	10	10	24.3000000	3.8600518	15.0000000	30.0000000
Alicante Bouschet	42	39	29.7179487	33.4474834	7.0000000	150.0000000
Aligoté	30	30	17.8333333	4.8358099	11.0000000	28.0000000
Alsace white blend	52	51	33.6470588	23.0649722	10.0000000	98.0000000
Altesse	1	1	18.0000000	.	18.0000000	18.0000000
Alvarelhão	2	2	18.0000000	0	18.0000000	18.0000000
Alvarinho	77	63	16.3492063	5.8672374	11.0000000	45.0000000
Alvarinho-Chardonn	3	2	10.0000000	1.4142136	9.0000000	11.0000000
Angevine	5	5	12.4000000	0.8944272	12.0000000	14.0000000

Analysis Variable : price						
variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Ansonica	4	1	18.0000000	.	18.0000000	18.0000000
Antão Vaz	16	15	23.4666667	5.3966480	13.0000000	30.0000000
Apple	6	6	31.0000000	4.7328638	25.0000000	35.0000000
Aragonez	9	8	24.1250000	14.2070154	10.0000000	45.0000000
Aragonês	15	13	30.5384615	21.9340503	8.0000000	70.0000000
Argaman	3	3	36.6666667	1.1547005	36.0000000	38.0000000
Arinto	72	54	16.1851852	6.9555844	7.0000000	40.0000000
Arneis	64	63	19.2857143	5.4252355	14.0000000	50.0000000
Asprinio	1	0	.	.	.	.
Assyrtico	67	67	23.3432836	6.4703338	13.0000000	40.0000000
Assyrtiko	8	8	21.5000000	4.8403070	17.0000000	30.0000000
Athiri	2	2	18.0000000	0	18.0000000	18.0000000
Austrian Red Blend	67	55	37.7636364	18.9882650	15.0000000	115.0000000
Austrian white ble	47	36	28.3888889	18.7102383	15.0000000	110.0000000
Auxerrois	17	14	24.6428571	4.4133912	16.0000000	32.0000000
Avesso	3	3	14.6666667	1.5275252	13.0000000	16.0000000
Azal	1	1	13.0000000	.	13.0000000	13.0000000
Baco Noir	9	9	24.2222222	4.2946996	18.0000000	30.0000000
Baga	22	16	31.6250000	22.4703508	9.0000000	70.0000000
Baga-Touriga Nacio	1	1	20.0000000	.	20.0000000	20.0000000

Analysis Variable : price						
variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Barbera	1365	967	25.9017580	14.2923363	9.0000000	163.0000000
Bastardo	7	7	30.5714286	0.9759001	30.0000000	32.0000000
Bical	13	9	15.2222222	7.7585079	9.0000000	28.0000000
Black Monukka	4	4	25.0000000	0	25.0000000	25.0000000
Black Muscat	13	13	25.9230769	9.1965713	10.0000000	38.0000000
Blatina	3	3	12.6666667	0.5773503	12.0000000	13.0000000
Blauburgunder	1	1	19.0000000	.	19.0000000	19.0000000
Blauer Portugieser	7	7	15.4285714	1.1338934	14.0000000	17.0000000
Blaufränkisch	227	191	29.0261780	16.8136644	9.0000000	129.0000000
Bobal	16	16	14.6875000	9.0753788	6.0000000	46.0000000
Bombino Bianco	1	1	30.0000000	.	30.0000000	30.0000000
Bonarda	152	152	15.0460526	5.3960236	9.0000000	38.0000000
Bordeaux-style Red	7347	4545	49.1634763	72.6755850	7.0000000	2300.00
Bordeaux-style Whi	1261	580	36.7206897	91.3422907	8.0000000	1000.00
Bovale	7	4	37.5000000	8.6602540	30.0000000	45.0000000
Boğazkere	3	3	25.0000000	6.9282032	21.0000000	33.0000000
Brachetto	25	24	18.2916667	4.0698164	11.0000000	27.0000000
Braucol	3	3	27.0000000	16.7032931	12.0000000	45.0000000
Bual	4	3	34.0000000	2.0000000	32.0000000	36.0000000
Bukettraube	2	2	18.0000000	0	18.0000000	18.0000000

Analysis Variable : price						
variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Cabernet	20	18	20.2222222	9.0719708	11.0000000	45.0000000
Cabernet Blend	305	301	61.0000000	59.6369572	8.0000000	500.0000000
Cabernet Franc	1363	1310	32.8152672	20.5014169	9.0000000	180.0000000
Cabernet Franc-Cab	3	3	34.0000000	6.9282032	26.0000000	38.0000000
Cabernet Franc-Car	6	6	18.5000000	17.9080987	10.0000000	55.0000000
Cabernet Franc-Mal	1	1	22.0000000	.	22.0000000	22.0000000
Cabernet Franc-Mer	10	9	45.5555556	17.6501495	28.0000000	80.0000000
Cabernet Franc-Tem	2	2	18.0000000	0	18.0000000	18.0000000
Cabernet Merlot	52	48	23.2083333	18.1412406	8.0000000	70.0000000
Cabernet Moravia	1	1	18.0000000	.	18.0000000	18.0000000
Cabernet Pfeffer	1	1	25.0000000	.	25.0000000	25.0000000
Cabernet Sauvignon	13470	13322	41.4960967	34.9645721	4.0000000	625.0000000
Cabernet-Shiraz	1	1	150.0000000	.	150.0000000	150.0000000
Cabernet-Syrah	12	12	26.0000000	7.9772404	16.0000000	40.0000000
Cannonau	43	35	35.2285714	22.3371041	15.0000000	91.0000000
Caprettone	1	1	19.0000000	.	19.0000000	19.0000000
Carignan	74	74	40.8378378	88.5230451	14.0000000	770.0000000
Carignan-Grenache	7	7	33.7142857	16.0801564	20.0000000	65.0000000
Carignan-Syrah	1	1	80.0000000	.	80.0000000	80.0000000
Carignane	26	25	25.1600000	6.9382995	11.0000000	42.0000000

Analysis Variable : price						
variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Carignano	66	58	38.9482759	21.2688904	11.0000000	91.0000000
Carineña	1	1	8.0000000	.	8.0000000	8.0000000
Cariñena-Garnacha	3	3	31.0000000	0	31.0000000	31.0000000
Carmenère	761	746	21.3270777	24.4216373	6.0000000	235.0000000
Carmenère-Caberne	22	20	16.0500000	2.3277502	13.0000000	20.0000000
Carmenère-Syrah	10	10	16.4000000	10.8852602	10.0000000	37.0000000
Carnelian	1	1	14.0000000	.	14.0000000	14.0000000
Carricante	23	22	44.5454545	37.3627358	21.0000000	195.0000000
Casavecchia	6	6	42.3333333	13.4709564	25.0000000	55.0000000
Castelão	37	37	10.8918919	2.5252479	7.0000000	17.0000000
Catalanesca	1	1	19.0000000	.	19.0000000	19.0000000
Catarratto	31	27	18.2962963	5.0825101	12.0000000	30.0000000
Cayuga	3	3	20.3333333	2.3094011	19.0000000	23.0000000
Cerceal	3	3	43.3333333	11.5470054	30.0000000	50.0000000
Cesanese d'Affile	18	9	22.0000000	7.5828754	16.0000000	35.0000000
Chambourcin	16	16	19.0000000	5.6920998	10.0000000	26.0000000
Champagne Blend	1238	1003	78.6271186	74.9159778	7.0000000	505.0000000
Charbono	40	40	31.3500000	6.1458762	16.0000000	40.0000000
Chardonel	1	1	11.0000000	.	11.0000000	11.0000000
Chardonelle	1	1	30.0000000	.	30.0000000	30.0000000

Analysis Variable : price						
variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Chardonnay	14482	13775	32.2471869	45.1487992	4.0000000	2013.00
Chardonnay Weissbu	3	3	25.0000000	0	25.0000000	25.0000000

SAS Code

```
/* Eric Fernandez Project-Phase B*/
/* I certify that the SAS code given is my original and exclusive work*/

/* To read the file:
```

```
 Create a  new folder.
 Upload 'winemag-data_first150k.csv' to the folder
 Right click on 'winemag-data_first150k.csv' and select Properties
 Copy the path name and paste to the filename statement below
 Add a slash and the file name to the end of the path
*/
FILENAME CSV "~/datasets/winemag-data_first150k.csv" TERMSTR=LF;


/** Import the CSV file.  **/
PROC IMPORT DATAFILE=CSV
                              OUT=WineReviews
                              DBMS=CSV
                              REPLACE;
RUN;

/* Section 2 */

/* 2(a) */
/* Single Categorical Variable: Country of Origin */
Proc gchart data=WineReviews;  /* general bar charting proc */
      pie country/type=percent; /* pie chart */
     title 'Countries Percentage' ;
Run;

/* 2(b) */
/* Single Quantitative Variable: Rating Points */
Proc sgplot data=WineReviews;
            histogram points;
            title 'Ratings';
Run;
title ;

/* 2(d) */
/* Created a new dataset with other countries that are not France, US,
  Spain or Italy merged into one group of countries called Other */
Data WineReviewsB;
            Set WineReviews;
            /* If countries are not Spain, US, Italy or France then change to Other*/
            if Country not in ('Spain' 'US' 'Italy' 'France') then country = 'Other';
Run;

/* Relationship between Quantitative and Categorial Response:
  Quantitative: Rating Points Categorical: Country of Origin*/
Proc sgplot data=WineReviewsB;
            vbox points / /* This is the quantitative variable for the y-axis */
            category = country; /* This is the categorical variable */
            title 'Relationship between Rating and Country of Origin';
Run;
title ;

/* 2(e) */
/* Relationship between Quantitative Variables:
  Quantitative Variables: x=Points y= Price*/
Proc sgplot data=WineReviews;
            scatter x=Points  y=Price; /* Quantitative Variables */
            title 'Relationship between Ratings and Prices';
Run;
```

```
title ;
/* Section 3 */

/* 3(a) */
/* Single Categorical Variable: Variety of Wine */
/* Counting varieties using proc freq */
Proc freq data=WineReviewsB;
        tables country; /* count the number of each type of variety */
        title 'Frequency of Country';
Run;
title ;

/* 3(b) */
/* Single Quantitative Variable: Price of Wines */
Proc means data=WineReviews;
        var Price;
        title 'Descriptive analysis of Prices of Wine';
Run;
title ;

/* 3(d) */
/* Relationship between Price and Variety  */
Proc means data=WineReviews;
   var price;
        class Variety;
        title 'Descriptive analysis of Prices of Wine per Variety';
Run;
title ;

/* 3(e) */
/* Relationship between points and price */
Proc corr data=WineReviews;
        var points price;
        title 'Relationship between Rating and Price';
Run;
title ;
```

STA3064

Regression Case Study

Background

In recent years, teen pregnancy has been a major issue in the United States. The objective of this study is to find out which economic indicators can be used to predict the teenage birth rate in large metropolitan areas. This study will focus on the city of Chicago. The hope is that by finding factors that contribute to teen pregnancy, a better understanding of prevention can be found to inform public policy makers.

Data Description and Variables

The data used comes from a larger data set containing 27 variables describing public health in various Chicago neighborhoods. The information was compiled by the Chicago Department of Public Health (CDPH). The accompanying file, *teen.csv*, contains eight variables and 77 observations. Designations of each variable follow. Variable names are given followed by a brief description in parentheses.

Predictor Variables:

- **BelowPovLev** (Below poverty level -- percent of households)
- **Crowded** (Crowded housing -- percent of occupied housing units)
- **Dependency** (Percent of people aged less than 16 or more than 64 years old)
- **NoHSDiploma** (No high school diploma -- percent of people aged 25 years or older)
- **Income** (Per capita income -- 2011 inflation-adjusted dollars)
- **Unemployment** (Percent of people not in labor force aged 16 years and older)

Response Variable:

- **BirthRate** (Teen birth rate -- per 1,000 females aged 15-19)

Tasks to Complete:

1. Data Exploration:

a. Keeping the text file external to your SAS code (i.e., do not use data lines), read your data into SAS. Include the data step used to get your data into a SAS data set. Print the first 20 observations. Comment on any additional data manipulation that was necessary.
b. Using your SAS data set, produce a scatterplot matrix of all variables (PROC SGSCATTER) and a correlation matrix (PROC CORR) of all variables and all observations. Note any interesting characteristics in the relationships revealed by the above procedures.

2. Model Fitting and Analysis:

From your analysis in the Data Exploration section above, assume that you select the variable, Unemployment, as the predictor variable that looks most promising in predicting your response. Use the following items to guide this portion of your analysis:
a. Fit a simple linear regression model. Provide the model equation.
b. Interpret the R2 for your fit.
c. Perform a residual analysis to determine if all model assumptions are being met.
d. Explore potential transformations on your response variable (even if your residual analysis indicates a transformation is not required just to confirm). If a transformation is indicated, refit your model and assess.
e. For your original simple regression model (2a) produce a 95% confidence interval for the true slope of your regression line. Interpret.

f. Create a 95% bootstrap confidence interval for the slope based on quantiles from the bootstrap distribution of at least 1000 replications. Compare your results to your confidence interval based on normal theory above.
g. Conduct the ANOVA test for the slope for the original in 2(a). Discuss whether this test indicates that the simple linear model is effective.
h. Suppose a new value of Unemployment of interest is 1.2% ($x^* = 1.2$). Produce both 95% confidence and prediction intervals around the predicted response for x^*. Interpret both intervals.

Now include all of your predictor variables in your analysis and use the following items to guide your study:

i. Explore the potential impact of multicollinearity on your full model using the original (non-transformed) data.
j. Use a variable selection method to assist in fitting your best multiple regression model (if you had to exclude any variables due to multicollinearity, do not include them in the variable selection procedure).
k. Interpret the R2 for your best model.
l. Perform a residual analysis to determine if all model assumptions are being met.
m. Perform a nested F-test comparing your full model (all predictors included) to a reduced model of interest.
n. Make note of any outliers, high-leverage points, and influential points. Describe how you think they may be impacting the fit of your model. Discuss whether removal of any points is justified.
o. At this point, state what you believe is your best model based on the above analysis. What would be your next step? (You do not need to perform the potential action, just discuss it.)

1) Data Exploration

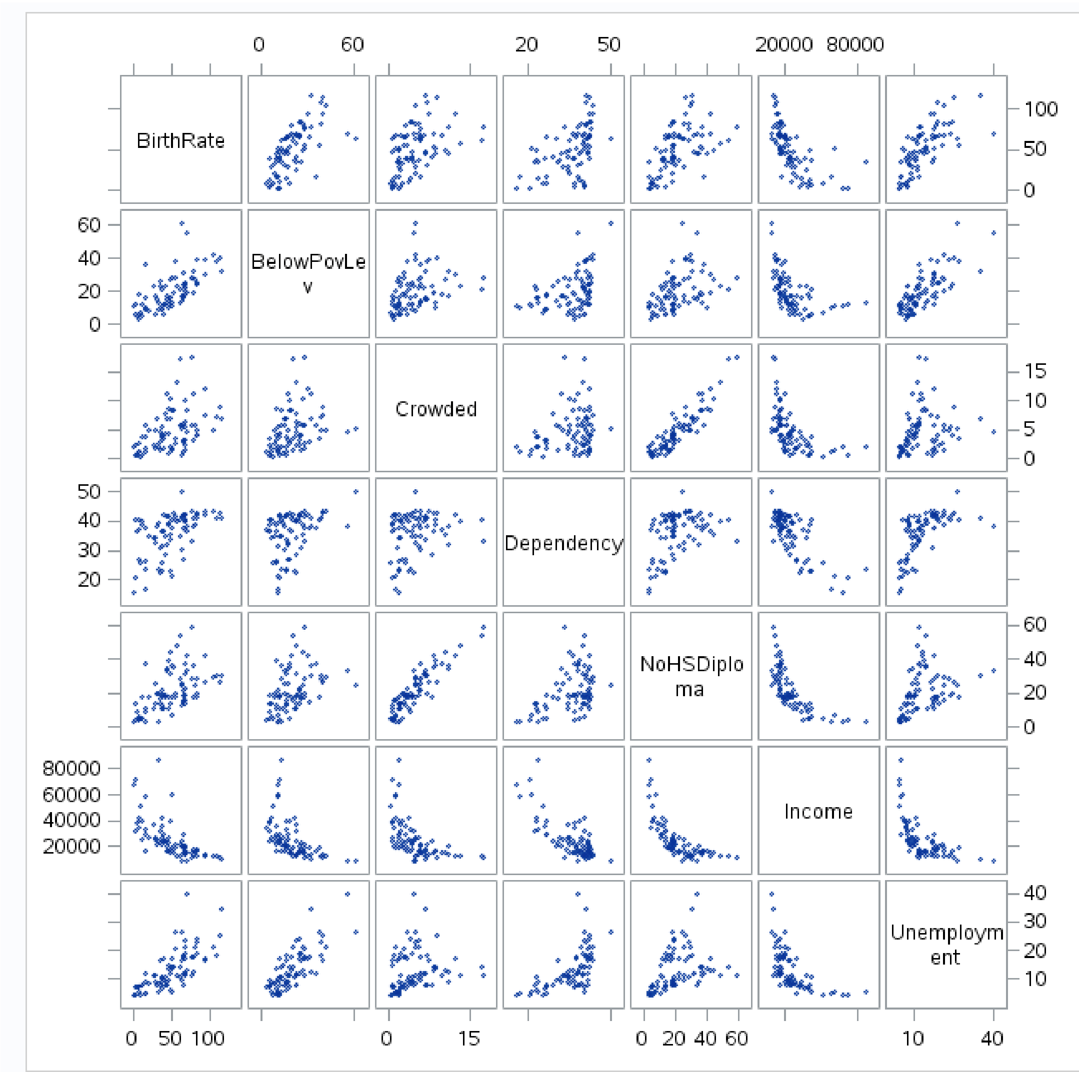
1(A)

First 20 observations:

Obs	Community	CommunityName	BirthRate	BelowPovLev	Crowded	Dependency	NoHSDiploma	Income	Unemployment
1	1	Rogers Park	40.8	22.7	7.9	28.8	18.1	23714	7.5
2	2	West Ridge	29.9	15.1	7	38.3	19.6	21375	7.9
3	3	Uptown	35.1	22.7	4.6	22.2	13.6	32355	7.7
4	4	Lincoln Square	38.4	9.5	3.1	25.6	12.5	35503	6.8
5	5	North Center	8.4	7.1	0.2	25.5	5.4	51615	4.5
6	6	Lake View	15.8	10.5	1.2	16.5	2.9	58227	4.7
7	7	Lincoln Park	2.1	11.8	0.6	20.4	4.3	71403	4.5
8	8	Near North Side	34	13.4	2	23.3	3.4	87163	5.2
9	9	Edison Park	3.9	5.1	0.6	36.6	8.5	38337	7.4
10	10	Norwood Park	3.4	5.9	2.3	40.6	13.5	31659	7.3
11	11	Jefferson Park	28.6	6.4	1.9	34.4	13.5	27280	9
12	12	Forest Glen	6.3	6.1	1.3	40.6	6.3	41509	5.5
13	13	North Park	10.5	12.4	3.8	39.7	18.2	24941	7.5
14	14	Albany Park	44.5	17.1	11.2	32.1	34.9	20355	9
15	15	Portage Park	41.7	12.3	4.4	34.6	18.7	23617	10.6
16	16	Irving Park	37	10.8	5.6	31.6	22	26713	10.3
17	17	Dunning	19.9	8.3	4.8	34.9	18	26347	8.6
18	18	Montclaire	61.5	12.8	5.8	35	28.4	21257	10.8
19	19	Belmont Cragin	68.2	18.6	10	36.9	37	15246	11.5
20	20	Hermosa	69.7	19.1	8.4	36.3	41.9	15411	12.9

1(B)

Scatterplot for all variables plotted against each other:



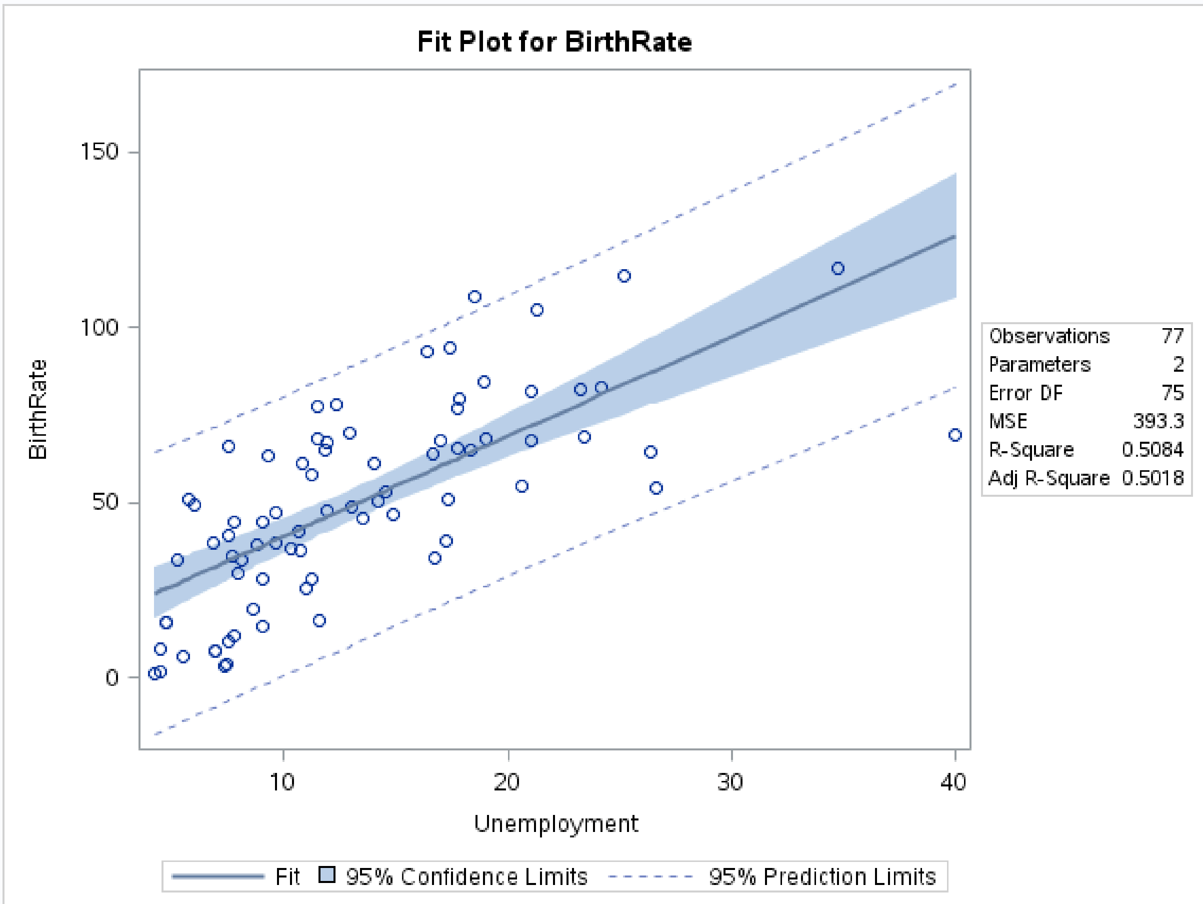
The scatter plots for **Crowded** vs **NoHSDiploma** show a positive linear relationship between variables.

Pearson Correlation Coefficients, N = 77 Prob > |r| under H0: Rho=0								
	Community	BirthRate	BelowPovLev	Crowded	Dependency	NoHSDiploma	Income	Unemployment
Community	1.00000	0.25016 0.0282	0.11026 0.3398	0.03461 0.7651	0.43210 <.0001	0.16127 0.1612	-0.36664 0.0010	0.33246 0.0031
BirthRate	0.25016 0.0282	1.00000	0.66004 <.0001	0.44840 <.0001	0.51788 <.0001	0.53778 <.0001	-0.64713 <.0001	0.71301 <.0001
BelowPovLev	0.11026 0.3398	0.66004 <.0001	1.00000	0.32324 0.0041	0.40135 0.0003	0.42238 0.0001	-0.52652 <.0001	0.76382 <.0001
Crowded	0.03461 0.7651	0.44840 <.0001	0.32324 0.0041	1.00000	0.24445 0.0321	0.90527 <.0001	-0.54520 <.0001	0.14430 0.2105
Dependency	0.43210 <.0001	0.51788 <.0001	0.40135 0.0003	0.24445 0.0321	1.00000	0.42436 0.0001	-0.75658 <.0001	0.60500 <.0001
NoHSDiploma	0.16127 0.1612	0.53778 <.0001	0.42238 0.0001	0.90527 <.0001	0.42436 0.0001	1.00000	-0.70735 <.0001	0.32290 0.0042
Income	-0.36664 0.0010	-0.64713 <.0001	-0.52652 <.0001	-0.54520 <.0001	-0.75658 <.0001	-0.70735 <.0001	1.00000	-0.61055 <.0001
Unemployment	0.33246 0.0031	0.71301 <.0001	0.76382 <.0001	0.14430 0.2105	0.60500 <.0001	0.32290 0.0042	-0.61055 <.0001	1.00000

The Pearson Correlation Coefficient chart above confirms the correlation between **Crowded** and **NoHSDiploma** by observing. It seem that the variables have a strong correlation (0.90527).

2) Model Fitting and Analysis

2(A)



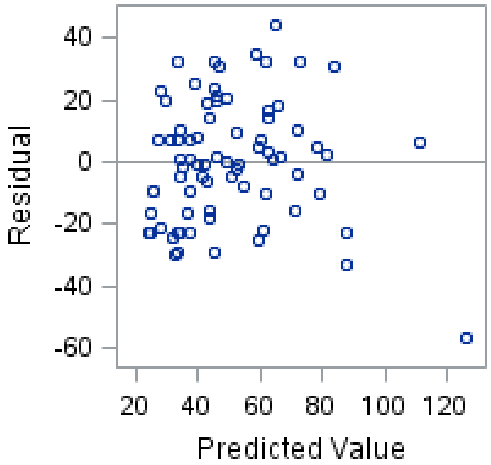
The equation for this model is

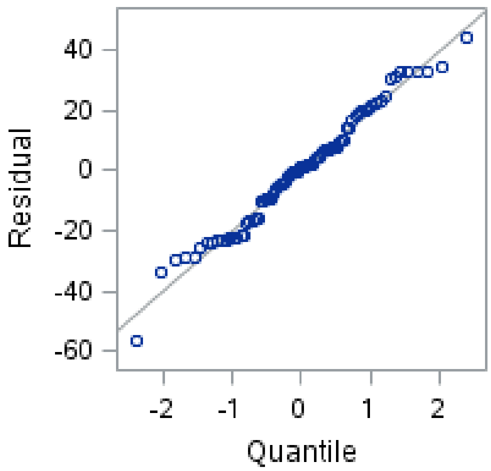
$$BirthRate = \beta_0 + \beta_1 Unemployment$$

2(B)

The R-Square is 0.5084 which means that **50.84%** of the variation of teenage female birth rates(**BirthRate** variable) can be explained by the percent of people not in the labor force aged 16 years and older(**Unemployment** variable).

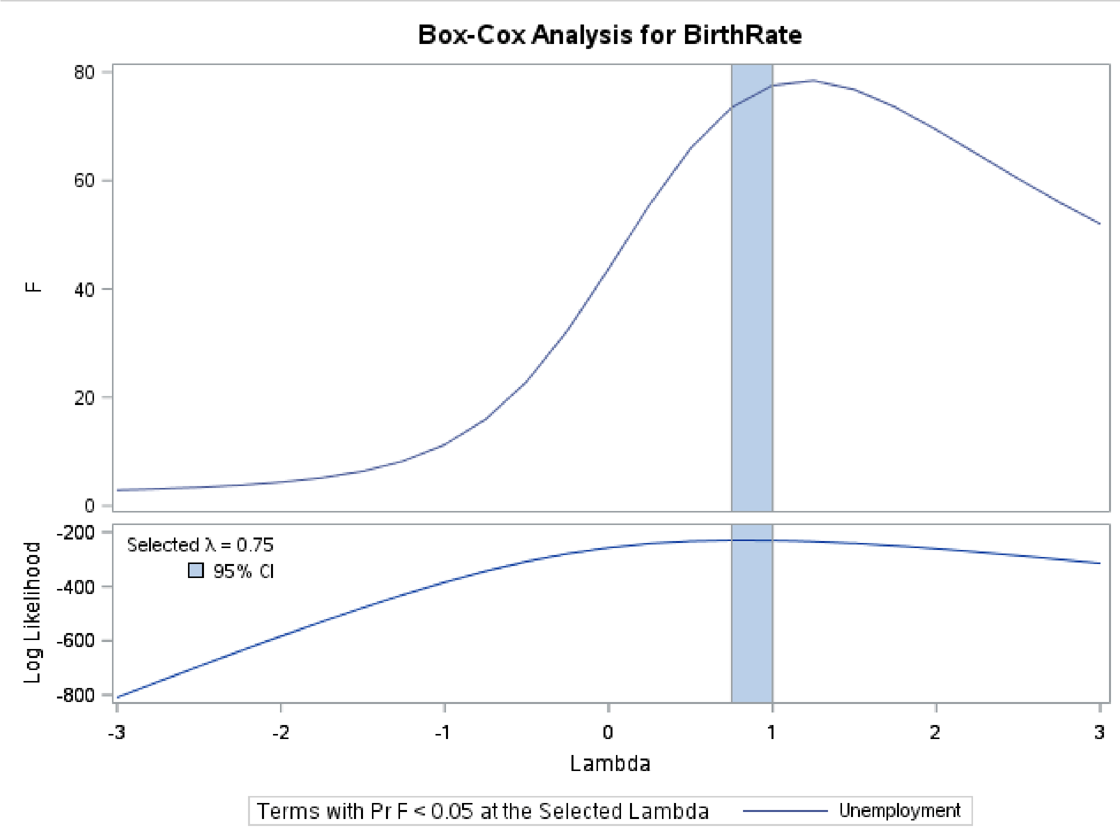
2(C)



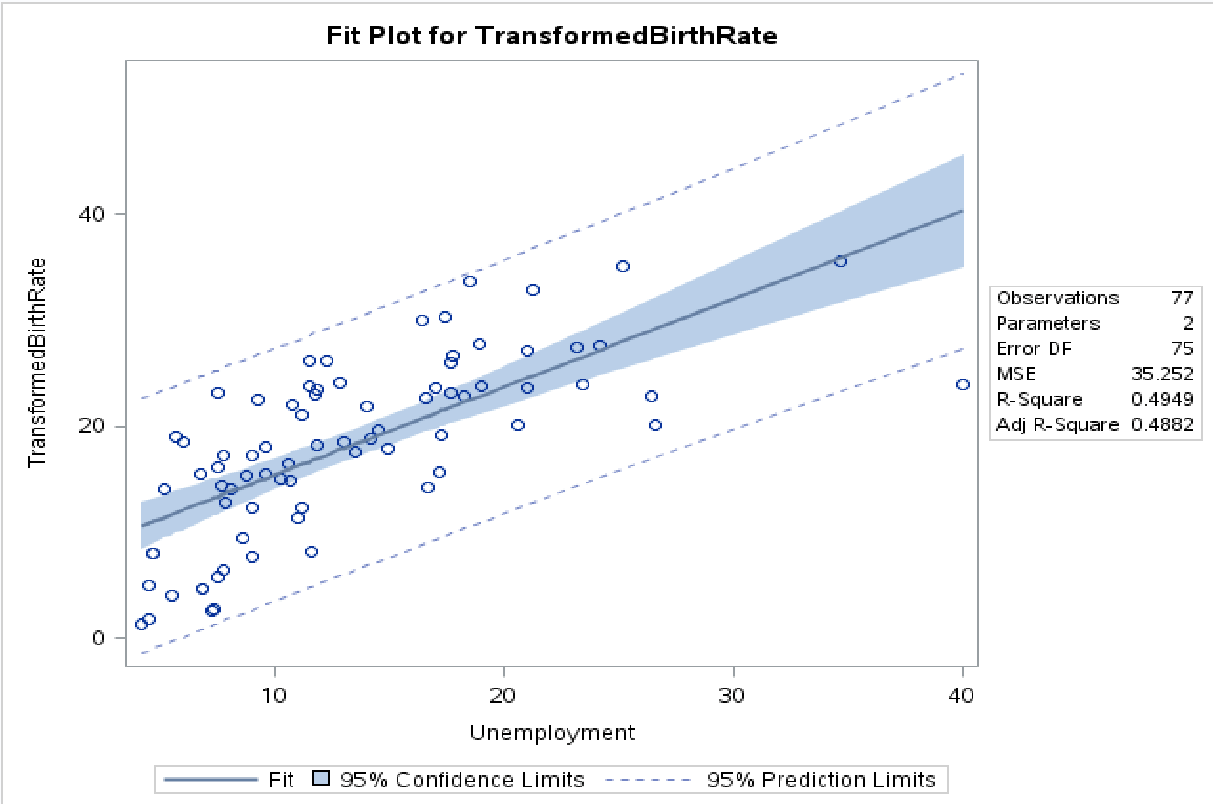


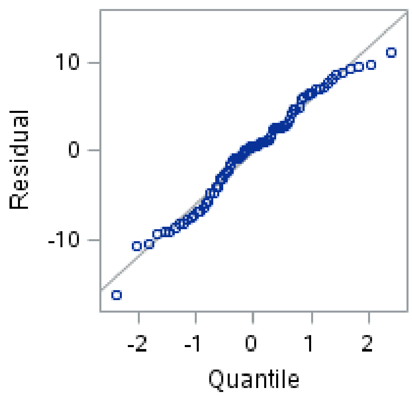
The graph of residual vs predicted values shows that the assumption of constant variance is met since the points are scattered randomly around 0 with what appears to be an outlier that has a predicted value above 120. The residual vs quantile plot graph show that the assumption of normality is also met.

2(D)



The Box-Cox method shows that the confidence Interval for the lambda value is between a number around 0.7 and 1. Having 1 in the interval indicates that this lambda value might not be useful for transformation since y^1=y. Nonetheless, I transformed the response using the value of lambda(0.75)





We can observe that R-Square has decreased and the residual vs quantile plot hasn't changed much indicating that this transformation is not be beneficial.

2(E)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	12.16201	4.86114	2.50	0.0145	2.47812	21.84589
Unemployment	1	2.84901	0.32350	8.81	<.0001	2.20456	3.49346

The 95% confidence of the interval for the slope is (2.20456,3.49346). This means that we are 95% confident that the population slope falls within the interval (2.20456, 3.49346).

2(F)

Bootstrap 95% Confidence Interval:

Obs	Conf_Limit_2_5	Conf_Limit_97_5
1	2.02783	3.85608

This shows that the bootstrap CI is narrower than the normal theory CI indicating that there is more variation on average for this model

2(G)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	62	56597.84366	912.86845	3.76	0.0043
Error	14	3403.19167	243.08512		
Corrected Total	76	60001.03532			

The ANOVA test for the model in 2(a) has a p-value of 0.0043. This indicates that at a 5% significance level, we can conclude that the simple linear model is effective.

2(H)

Output Statistics									
Obs	Unemployment	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	1.2	.	15.5808	4.5210	6.5744	24.5872	-24.9395	56.1012	.

The 95% confidence interval for the mean birthrate of female teenagers with 1.2 percent of unemployment is (6.5744,24.5872) and the 95% prediction interval is (-24.9395,56,1012). The width for the 95% confidence interval for the mean birthrate teenage females with 1.2 percent of unemployment is smaller than the width for the 95% confidence interval for a particular birthrate of a teenage female with 1.2 percent of unemployment because there is more variation in the a particular female teenager than if you take the mean of the female teenagers.

2(I)

Full Model VIF:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	15.43827	24.09916	0.64	0.5239	0
BelowPovLev	1	0.30175	0.29304	1.03	0.3067	2.81644
Crowded	1	2.47483	1.43586	1.72	0.0892	6.84296
Dependency	1	0.01495	0.46114	0.03	0.9742	2.78867
NoHSDiploma	1	-0.17453	0.48199	-0.36	0.7184	8.79952
Income	1	-0.00028394	0.00028511	-1.00	0.3227	4.50964
Unemployment	1	2.00747	0.54892	3.66	0.0005	3.69714

The estimates show that **Crowded** and **NoHSDiploma** have a variance inflation factor above 5. **NoHSDiploma** has the highest variance inflation factor so I decided to remove this variable from the model since it is a cause for multicollinearity and tried a regression model without **NoHSDiploma.**

Model Without NoHSDiploma VIF:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	13.40971	23.29509	0.58	0.5667	0
BelowPovLev	1	0.30134	0.29124	1.03	0.3043	2.81640
Crowded	1	2.02828	0.73089	2.78	0.0070	1.79506
Dependency	1	0.00822	0.45794	0.02	0.9857	2.78415
Income	1	-0.00024828	0.00026592	-0.93	0.3536	3.97163
Unemployment	1	1.99298	0.54410	3.66	0.0005	3.67748

The variance inflation factor of all variables after removing **NoHSDiploma** are all below 5. No other variable will be removed from the model in this step.

2(J)

Stepwise Selection: Step 2

Variable Crowded Entered: R-Square = 0.6303 and C(p) = 2.3211

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	37819	18909	63.08	<.0001
Error	74	22182	299.75893		
Corrected Total	76	60001			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.55043	4.75658	31.84847	0.11	0.7454
Crowded	2.71085	0.54876	7315.01372	24.40	<.0001
Unemployment	2.64555	0.28541	25755	85.92	<.0001

Bounds on condition number: 1.0213, 4.0851

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Unemployment		1	0.5084	0.5084	24.5002	77.56	<.0001
2	Crowded		2	0.1219	0.6303	2.3211	24.40	<.0001

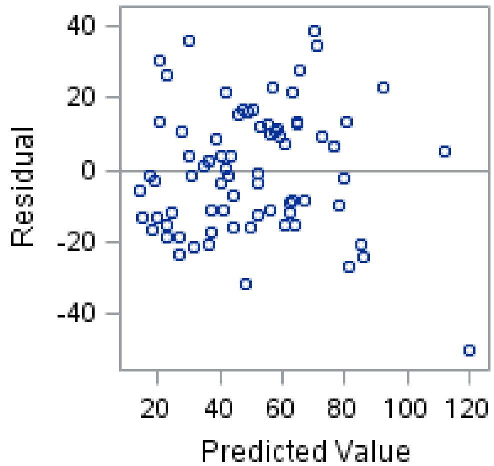
Using the stepwise selection method, the charts show that the only significant variables are **Unemployment** and **Crowded**.

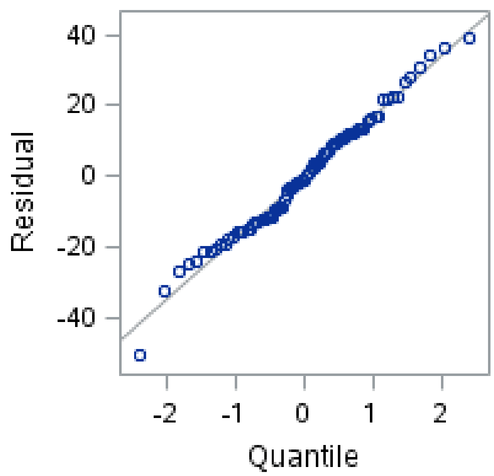
2(K)

Root MSE	17.31355	R-Square	0.6303
Dependent Mean	50.06494	Adj R-Sq	0.6203
Coeff Var	34.58218		

The R-Square of the suggested best model in step 2(j) is 0.6303 which means that **63.03%** of the variation of teen birthrates(**BirthRate** variable) can be explained by the percent of people not in labor force aged 16 years and older(**Unemployment** variable), and the percent of occupied housing units(**Crowded** variable).

2(L)





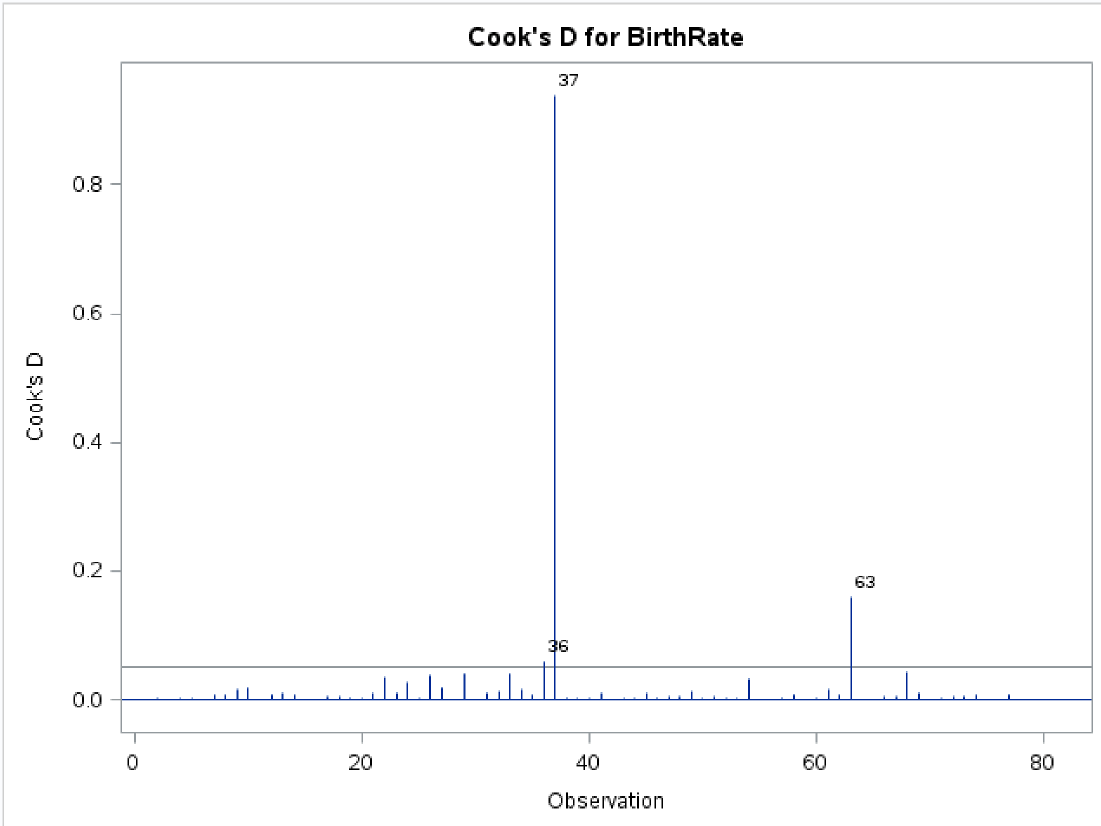
The graph of residual vs predicted values shows that the assumption of constant variance is met since the points are scattered randomly around 0 with what appears to be an outlier that has predicted value above 120. The residual vs quantile plot graph display has improved from the model from 2(a), showing very light tails. This means the best model meets the normality assumption.

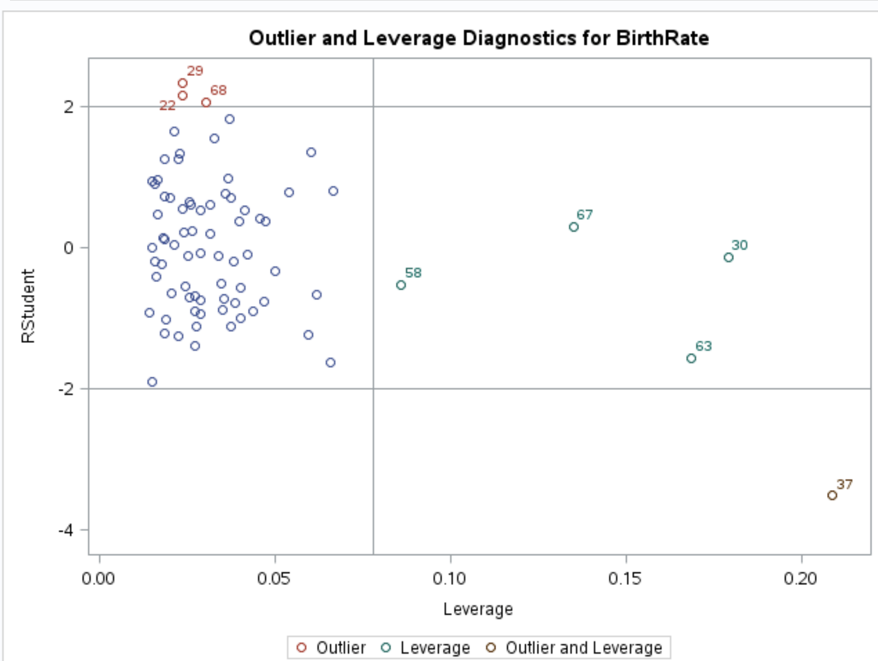
2(M)

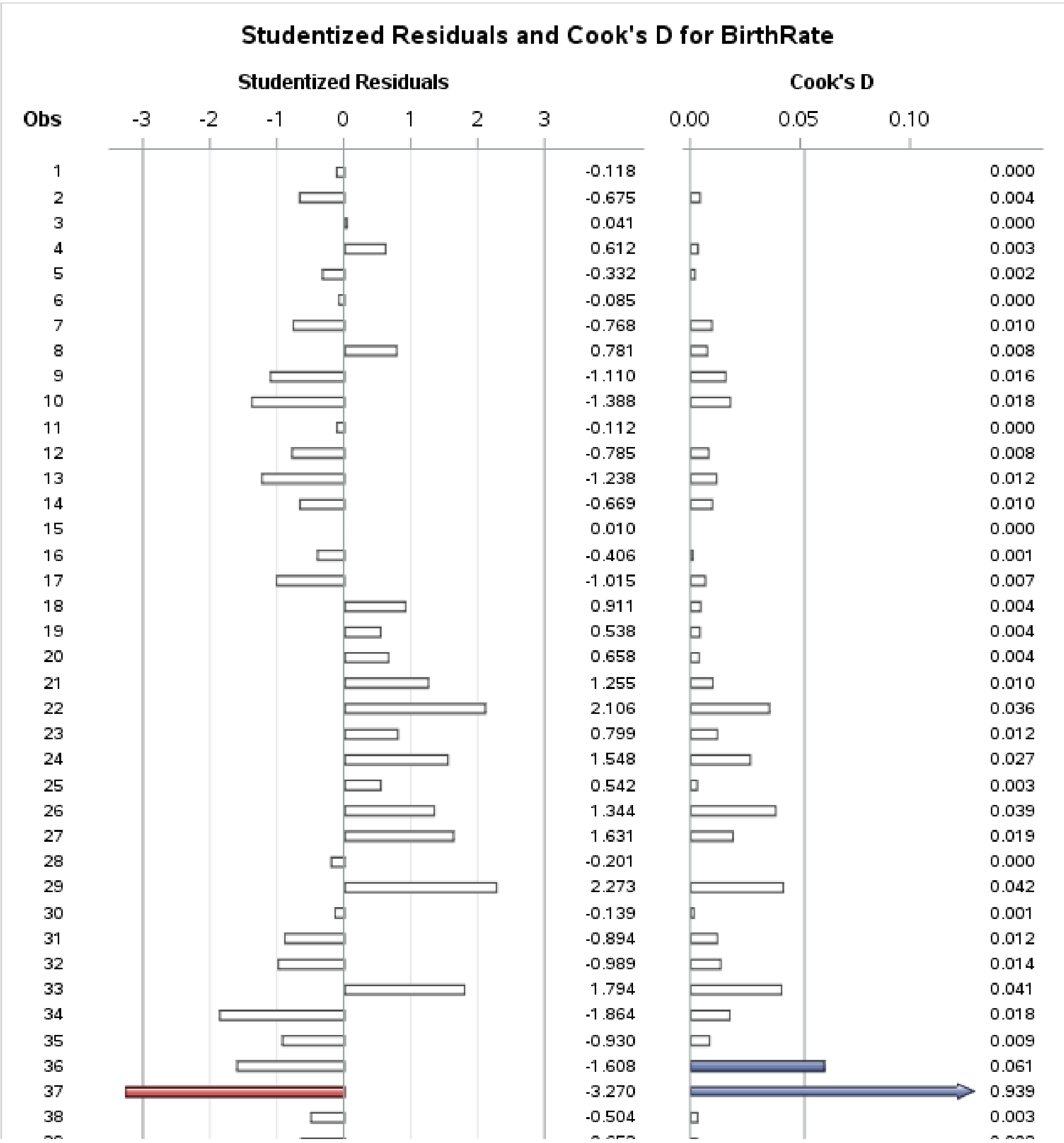
Test 1 Results for Dependent Variable BirthRate				
Source	DF	Mean Square	F Value	Pr > F
Numerator	3	234.07375	0.77	0.5125
Denominator	71	302.53436		

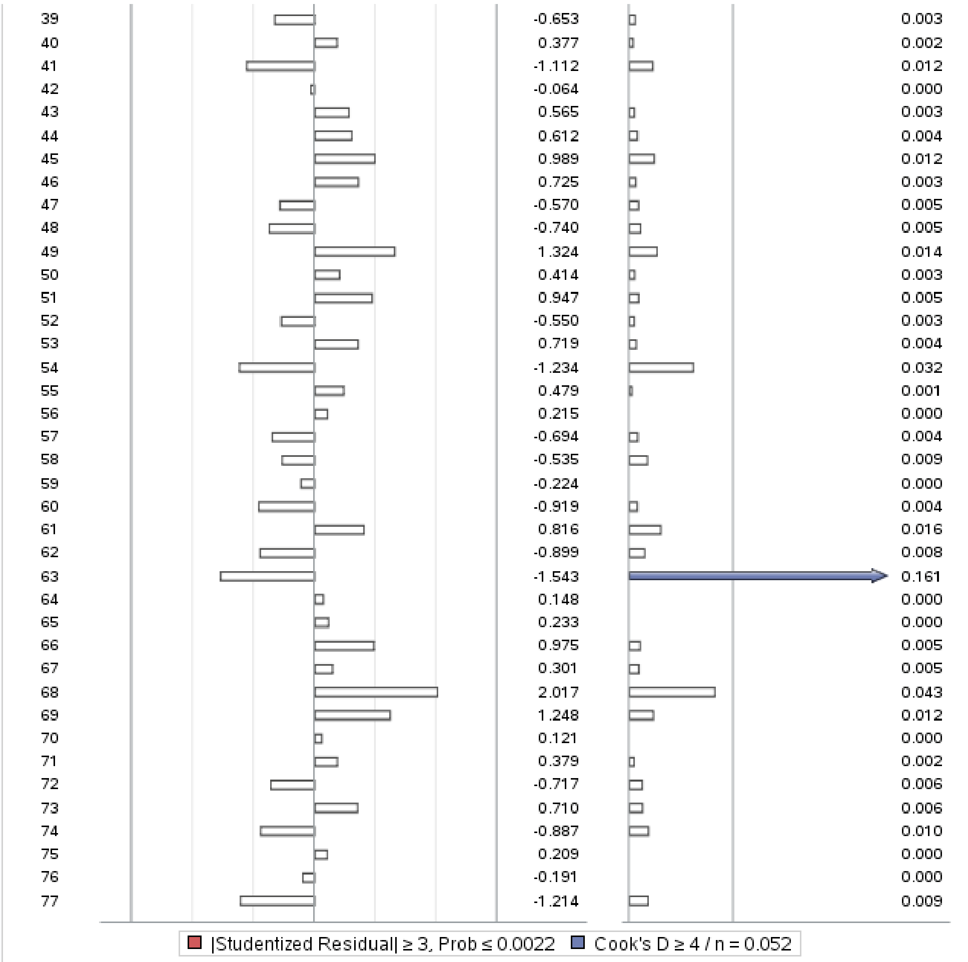
The nested f-test has a p-value of 0.5125 meaning the variables **BelowPovLevel**, **Dependency** and **Income** of the reduced model are not significant at a 5% significance level when predicting **BirthRate**.

2N)


Cook's D for BirthRate
37
63
36
0.8
0.6
0.4
0.2
0.0
Cook's D
0
20
40
60
80
Observation


Outlier and Leverage Diagnostics for BirthRate
29
68
22
67
30
58
63
37
2
0
-2
-4
RStudent
0.00
0.05
0.10
0.15
0.20
Leverage
Outlier
Leverage
Outlier and Leverage


Studentized Residuals and Cook's D for BirthRate
Studentized Residuals
Cook's D
Obs
-3
-2
-1
0
1
2
3
0.00
0.05
0.10
1 -0.118 0.000
2 -0.675 0.004
3 0.041 0.000
4 0.612 0.003
5 -0.332 0.002
6 -0.085 0.000
7 -0.768 0.010
8 0.781 0.008
9 -1.110 0.016
10 -1.388 0.018
11 -0.112 0.000
12 -0.785 0.008
13 -1.238 0.012
14 -0.669 0.010
15 0.010 0.000
16 -0.406 0.001
17 -1.015 0.007
18 0.911 0.004
19 0.538 0.004
20 0.658 0.004
21 1.255 0.010
22 2.106 0.036
23 0.799 0.012
24 1.548 0.027
25 0.542 0.003
26 1.344 0.039
27 1.631 0.019
28 -0.201 0.000
29 2.273 0.042
30 -0.139 0.001
31 -0.894 0.012
32 -0.989 0.014
33 1.794 0.041
34 -1.864 0.018
35 -0.930 0.009
36 -1.608 0.061
37 -3.270 0.939
38 -0.504 0.003



By looking at these charts, we can observe that there are 3 observations with higher than normal Cook's distance values(36,37,38) with one been almost 1(37) , 5 with high leverage(58,67,30, 63,37)and one with a very low R-Student value(37). Observations with a large R-Student values(in magnitude) indicate unusual response values, observations with high leverage indicate covariate values that are extreme(far from the center of the distribution) and observations with a high Cook's D indicate values that a high influence on the estimated parameters and the predicted values. I decided to remove observation 37 since it is highly influential, has a high leverage and a very low R-Student Value.

By removing the influential observation, we can see that the R-Square improved by 5%.

Root MSE	16.12305	**R-Square**	0.6818
Dependent Mean	49.81316	**Adj R-Sq**	0.6730
Coeff Var	32.36704		

This indicates that now the predictors explain more in the variation of the response.

2(O)

I would suggest using the best model without the influential/outlier observation 37 in order to improve the prediction.

SAS CODE

```
/* Eric Fernandez Case Study */
/* I certify that the SAS code given is my original and exclusive work */
/* Part 1 DATA EXPLORATION */
/* 1(a) */
/* Datastep */
/*
 * To read the file:
 * -Locate the file 'Teen.csv'
 * -Right click on 'Teen.csv' and select 'Properties'
 * -Copy and paste the path name to the FILENAME statement
 */
FILENAME CSV "/home/eff100/datasets/Teen.csv" TERMSTR=CRLF;
PROC IMPORT DATAFILE=CSV
                    OUT=teen
                    DBMS=CSV
                    REPLACE;
RUN;
/* Print first 20 observations */
PROC PRINT data=teen(obs=20);
RUN;

/* 1(b) */
/* Scatter plot for all the variables plotted against each other */
PROC SGSCATTER data=teen;
        matrix BirthRate BelowPovLev Crowded Dependency NoHSDiploma Income Unemployment;
RUN;

/* Output correlation coefficient table */
PROC CORR data=teen;
RUN;

/* Part 2 MODEL FITTING AND ANALYSIS */
/* 2(a), 2(b), 2(c) and 2(d) */
/* Look at ANOVA test and R-square*/
PROC REG data=teen;
        model BirthRate = Unemployment;
RUN;



/* Perform Box-Cox test to obtain lambda for power transformation */
PROC TRANSREG data=teen;
        model Boxcox(BirthRate)=Identity(Unemployment);
RUN;

/* Transform response using the lambda obtained from Box-Cox */
DATA TeenTransformed;
        set teen;
        TransformedBirthRate=BirthRate**0.75;
RUN;
/*
 * Look at R2 and residual vs quantile plot to asses if transformation
 * is beneficial
 */
PROC REG data=TeenTransformed;
```

```
        model TransformedBirthRate = Unemployment;
RUN;

/* 2(e) */
/* Confidence intervals -- 95% default */
PROC REG data=teen;
        model BirthRate = Unemployment/clb; /* Confidence interval for the slope */
RUN;

/* 2(f) */
/* Bootstrap for slope Confidence Interval */
/* Generate 1000 replications with equal probability and with replacement(URS). */
PROC SURVEYSELECT Data=teen out=boot
        seed=4321 samprate=1
        method=urs outhits rep=1000;
RUN;
/* Regression on every sample to find the slopes */
PROC REG data=boot outest=betas noprint;
        model BirthRate= Unemployment;
        by replicate;
RUN;
PROC UNIVARIATE data=betas noprint;
        var Unemployment;
        output out=BootCI pctlpts= 2.5 97.5  pctlpre=Conf_Limit_;
RUN;
/* Print 95% Confidence interval of the bootstrap set */
PROC PRINT data=BootCI;
RUN;

/* 2(g) */
/* Perform ANOVA on the model of 2(a) to verify model's effectiveness */
PROC ANOVA data=teen;
        class Unemployment;
        model BirthRate = Unemployment;
RUN;

/* 2(h) */
/* Create dataset with Unemployment=1.2 */
DATA NewTeen;
Input BirthRate Unemployment;
datalines;
. 1.2
;
RUN;

/* Create new Data set with old values plus the value created. */
DATA Teen2;
        set NewTeen teen; /* Concatenate Data Sets */
RUN;

/* Produce both 95% confidence and prediction intervals around the predicted response for x*.  */
PROC REG data=Teen2;
        model BirthRate = Unemployment/cli clm;
        id unemployment;
RUN;

/* 2(i) */
PROC REG data=Teen;
```

```
	model BirthRate = BelowPovLev Crowded Dependency NoHSDiploma Income Unemployment /VIF; /* VIF checks for multicolinearity */
RUN;
/* Model with no NoHSDiploma due to high VIF*/
PROC REG data=Teen;
	model BirthRate = BelowPovLev Crowded Dependency Income Unemployment /VIF; /* VIF checks for multicolinearity */
RUN;

/* 2(j) */
/* Stepwise method for variable selection. This model does not include NoHSDiploma*/
PROC REG data=Teen;
	model BirthRate = BelowPovLev Crowded Dependency Income Unemployment/selection=stepwise;
RUN;

/* 2(k), 2(l) and 2(n) */
/* Do a regression using only the variables obtained from the stepwise selection method */
PROC REG data=Teen plots(label)=(CooksD RStudentByLeverage);
	model BirthRate = Crowded Unemployment/r influence;
RUN;

data TeenNew ;
 set teen;
 if community = 37 then delete;/* remove obs 37*/
RUN;

/* fit model with new dataset */
PROC REG data=TeenNew;
	model BirthRate = Crowded Unemployment/r influence;
RUN;

/* 2(m) */
/*Nested F test*/
PROC REG data=Teen;
	model BirthRate = BelowPovLev Crowded Dependency Income Unemployment;
	test BelowPovLev,Dependency,Income ;
RUN;
```

STA4203

Assignment 6

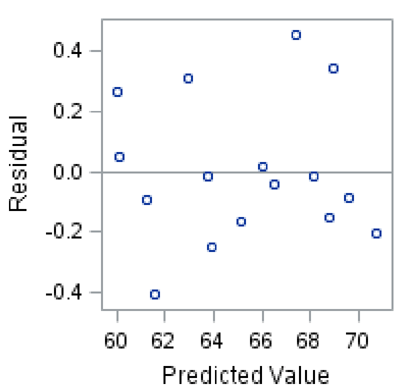
PROBLEM 1)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	184.17240	30.69540	330.29	<.0001
Error	9	0.83642	0.09294		
Corrected Total	15	185.00883			

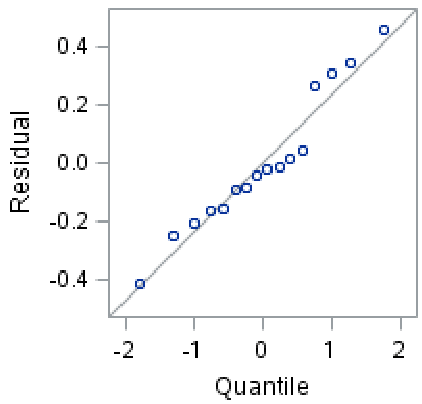
Root MSE	0.30485	R-Square	0.9955
Dependent Mean	65.31700	Adj R-Sq	0.9925
Coeff Var	0.46673		

Using the model with all 6 predictors, we observe that **99.55 percent** of the variance in the number of people employed can be explained by the GNP deflator, GNP, number of unemployed, number of people in the armed forces, the 'noninstitutionalized' population with more than 14 years of age and the year. At a 5% significance level, we conclude that there is a significant relationship between the predictors and the number of people employed.

PROBLEM 2)

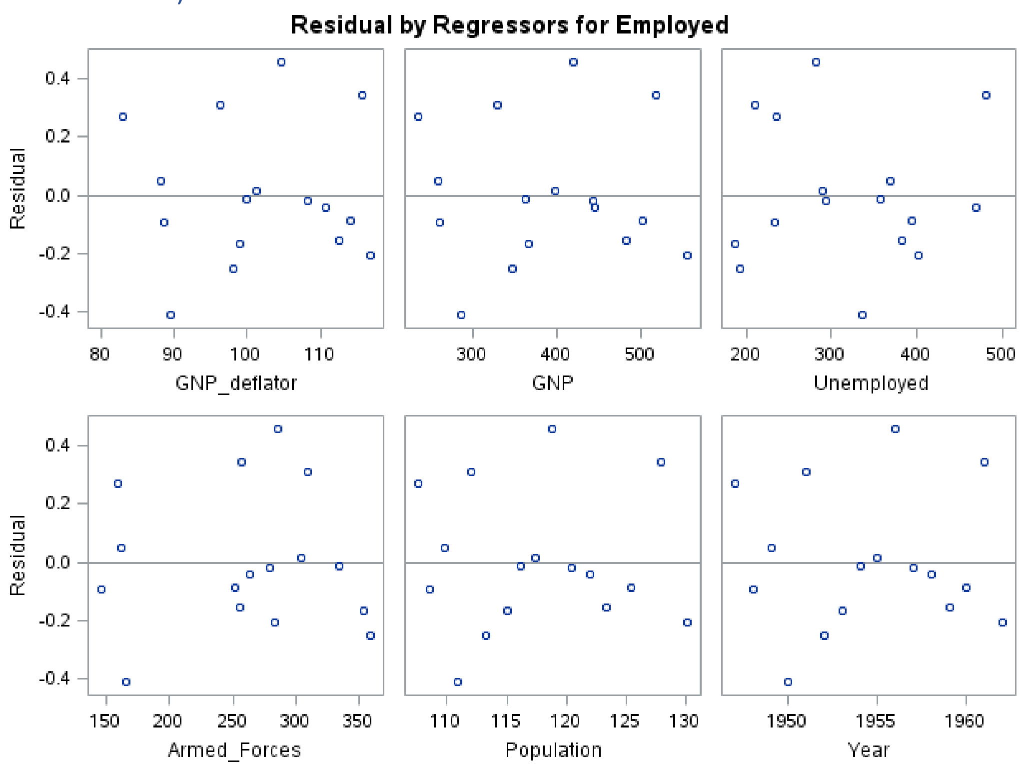


The residuals vs predicted values plot indicates that the assumption of constant variance is met since the points are scattered randomly around 0.



Analyzing the residual vs quantile plot, we can observe that the assumption of normality.

PROBLEM 3)



Since the residuals vs regressors plots don't seem to follow a pattern we can assume homoscedasticity. However, the standard error values varies across independent variables meaning that the error terms could be dependent.

PROBLEM 4)

Correlation							
Variable	GNP_deflator	GNP	Unemployed	Armed_Forces	Population	Year	Employed
GNP_deflator	1.0000	0.9916	0.6206	0.4647	0.9792	0.9911	0.9709
GNP	0.9916	1.0000	0.6043	0.4464	0.9911	0.9953	0.9836
Unemployed	0.6206	0.6043	1.0000	-0.1774	0.6866	0.6683	0.5025
Armed_Forces	0.4647	0.4464	-0.1774	1.0000	0.3644	0.4172	0.4573
Population	0.9792	0.9911	0.6866	0.3644	1.0000	0.9940	0.9604
Year	0.9911	0.9953	0.6683	0.4172	0.9940	1.0000	0.9713
Employed	0.9709	0.9836	0.5025	0.4573	0.9604	0.9713	1.0000

From the correlation table, we can observe that the predictors **GNP_deflator-GNP, GNP_deflator-Unemployed, GNP_deflator-Population, GNP_deflator-Year, GNP-Unemployed, GNP-Population, GNP-Year, Unemployed-Population, Unemployed-Year and Population-Year** are highly correlated pairs of predictors because their values are all greater than 0.6. Since most of the variables are correlated, I would say this dataset has a serious problem with multicollinearity.

PROBLEM 5)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-3482.25864	890.42038	-3.91	0.0036	0
GNP_deflator	1	0.01506	0.08491	0.18	0.8631	135.53244
GNP	1	-0.03582	0.03349	-1.07	0.3127	1788.51349
Unemployed	1	-0.02020	0.00488	-4.14	0.0025	33.61889
Armed_Forces	1	-0.01033	0.00214	-4.82	0.0009	3.58893
Population	1	-0.05110	0.22607	-0.23	0.8262	399.15102
Year	1	1.82915	0.45548	4.02	0.0030	758.98060

The VIF for **GNP_deflator, GNP, Unemployed, Population** and **Year** are all greater than 10. This indicates that the estimates for these predictors are highly inflated by multicollinearity.

PROBLEM 6)

The only predictor not dependent on others is **Armed_Forces**. This means it is the only predictor that is orthogonal to the others. The R2 values for the predictors are all above the 0.3 threshold, **Armed_Forces** being the lowest at 0.67.

PROBLEM 7)

Collinearity Diagnostics									
		Condition	Proportion of Variation						
Number	Eigenvalue	Index	Intercept	GNP_deflator	GNP	Unemployed	Armed_Forces	Population	Year
1	6.86139	1.00000	1.54013E-10	0.00000164	6.742617E-7	0.00004472	0.00035369	1.740763E-7	1.54148E-10
2	0.08210	9.14172	8.16629E-10	7.095535E-9	0.00000753	0.01428	0.09191	4.021693E-8	7.70535E-10
3	0.04568	12.25574	3.342247E-8	1.012272E-7	0.00025717	0.00083626	0.06357	0.00000839	3.19652E-8
4	0.01069	25.33661	1.19104E-9	0.00034484	0.00107	0.06464	0.42672	0.00001821	1.426706E-9
5	0.00012923	230.42395	5.260203E-7	0.45677	0.01566	0.00559	0.11540	0.00968	5.273968E-7
6	0.00000625	1048.08030	0.00014914	0.50456	0.32839	0.22534	6.865016E-7	0.83056	0.00016031
7	3.663846E-9	43275	0.99985	0.03833	0.65463	0.68926	0.30205	0.15973	0.99984

There are three condition indices which are greater than 100. This indicates that there is **strong** collinearity.

PROBLEM 8)

The collinearity diagnostic above shows three condition indices greater than 100. This indicates that there might be 3 strong sources causing multicollinearity.

SAS CODE

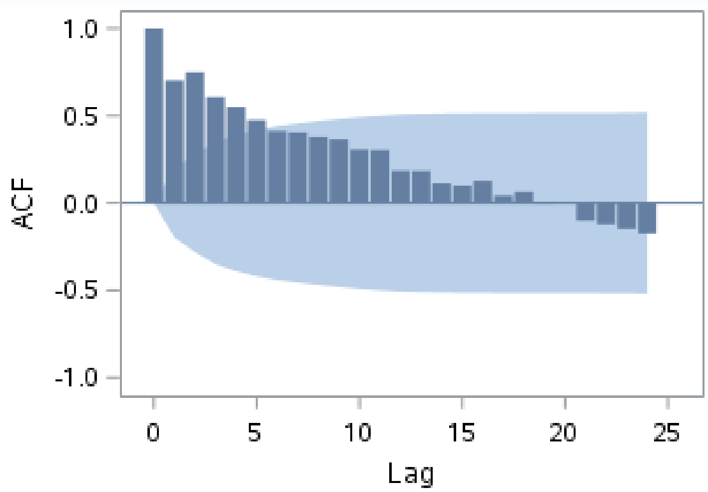
```
/* Read file spider.txt and store it in dataset spiders */
FILENAME longley '/home/eff100/my_courses/jhshows0/Data Sets/longley.txt';
Data macroecon;
INFILE longley;
INPUT GNP_deflator GNP Unemployed Armed_Forces Population Year Employed;
run;

/* Problem 1-5*/
/*SSR for full model*/
PROC REG data=macroecon;
MODEL Employed= GNP_deflator GNP Unemployed Armed_Forces Population Year/vif;
plot r.*p.;
OUTPUT out=resids1 r=resid p=pred;
run;

/* Check for normality */
PROC UNIVARIATE data=resids1 normal plots;
var resid;
run;

/* Look for correlation between variables by using coefficients */
proc corr data=macroecon;
var GNP_deflator GNP Unemployed Armed_Forces Population Year;
run;
```

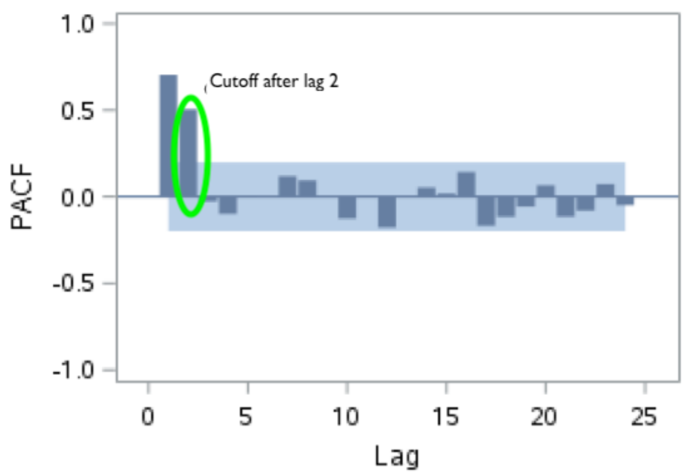
STA4853

Homework 2

PROBLEM 1)

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	225.32	6	<.0001	0.703	0.750	0.607	0.549	0.476	0.413
12	299.96	12	<.0001	0.406	0.379	0.364	0.307	0.303	0.185
18	309.40	18	<.0001	0.184	0.114	0.101	0.126	0.043	0.064
24	319.78	24	<.0001	-0.006	0.002	-0.101	-0.122	-0.148	-0.176

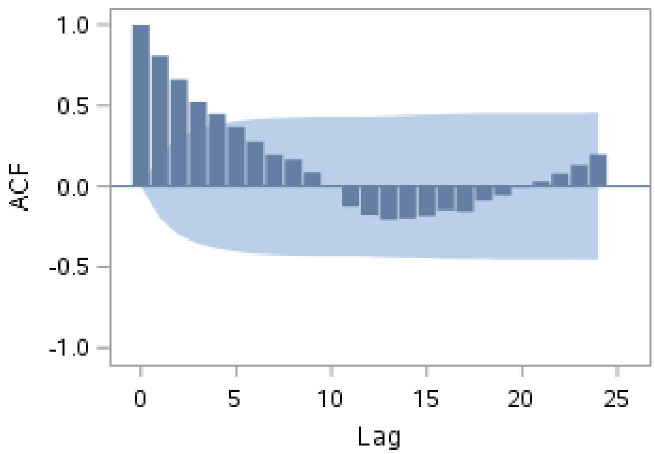


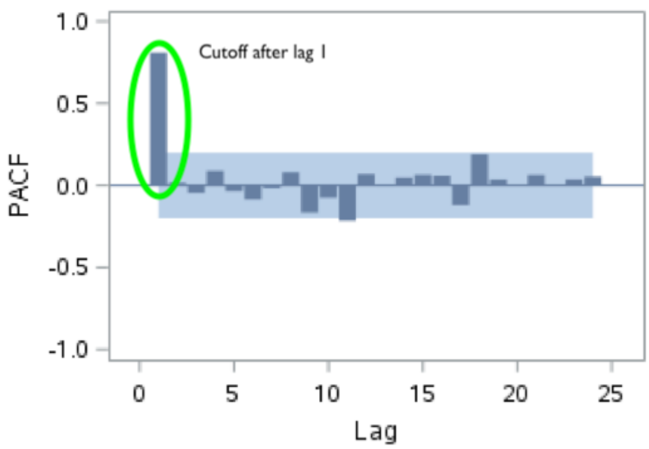


This series was generated by an **AR(2)** process because the ACF seems to decay exponentially to zero and the PACF has a cutoff after lag 2.

PROBLEM 2)

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	187.68	6	<.0001	0.810	0.661	0.525	0.447	0.367	0.275
12	201.34	12	<.0001	0.196	0.164	0.085	0.007	-0.127	-0.178
18	222.08	18	<.0001	-0.208	-0.201	-0.184	-0.148	-0.155	-0.089
24	231.02	24	<.0001	-0.053	-0.007	0.031	0.078	0.133	0.198

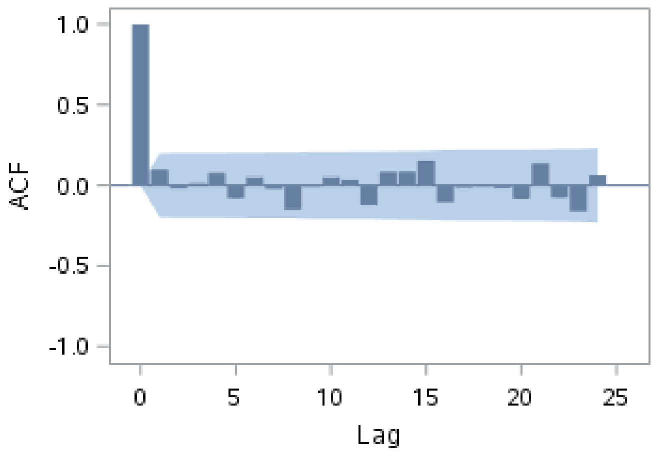


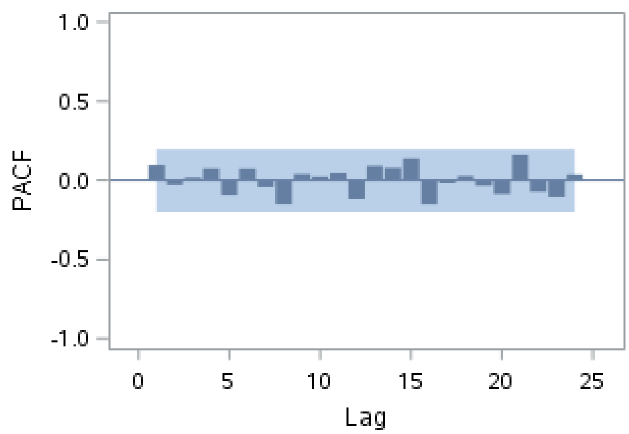


This series was generated by an **AR(1)** process because the ACF seems to decay exponentially to zero and the PACF has a cutoff after lag 1.

PROBLEM 3)

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	2.66	6	0.8498	0.097	-0.019	0.012	0.078	-0.080	0.051
12	7.46	12	0.8260	-0.023	-0.149	-0.008	0.051	0.035	-0.125
18	13.38	18	0.7684	0.083	0.084	0.154	-0.105	-0.013	-0.008
24	21.44	24	0.6126	-0.017	-0.084	0.137	-0.076	-0.160	0.061

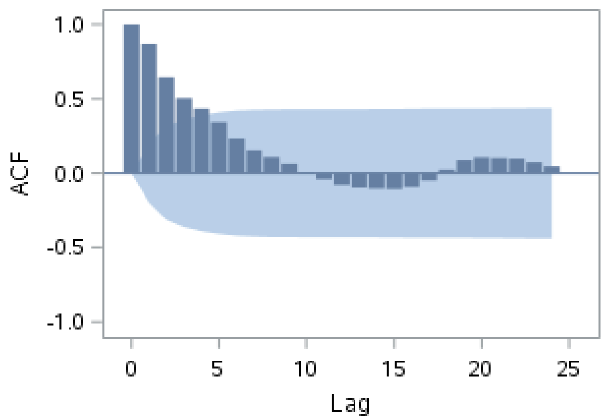


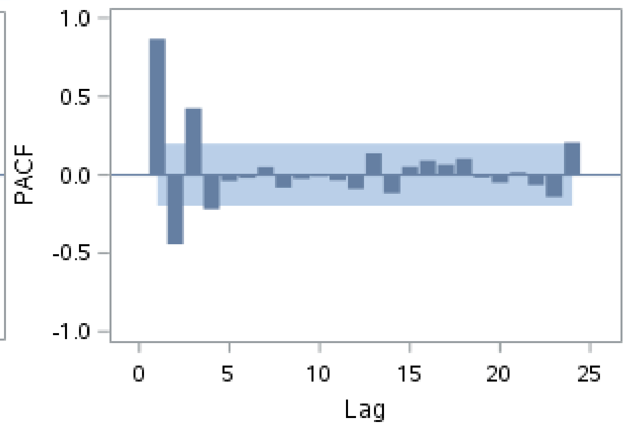


This series was generated by **random shocks**. By looking at the PACF and ACF, it is difficult to define which process was used. However, when we check the Autocorrelation Check for White Noise chart, we can observe that all the p-values are above 0.05 meaning that all this series was generated by random shocks.

PROBLEM 4)

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	188.16	6	<.0001	0.868	0.644	0.506	0.434	0.344	0.233
12	193.48	12	<.0001	0.154	0.110	0.062	0.005	-0.040	-0.078
18	198.69	18	<.0001	-0.098	-0.104	-0.107	-0.093	-0.048	0.023
24	204.85	24	<.0001	0.087	0.108	0.104	0.100	0.073	0.046

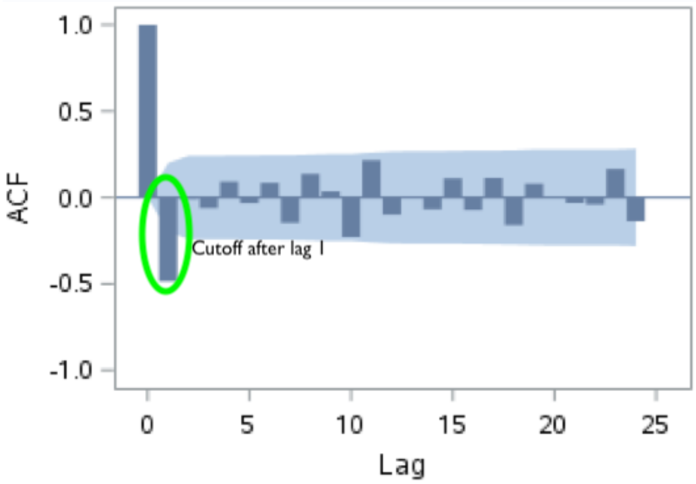


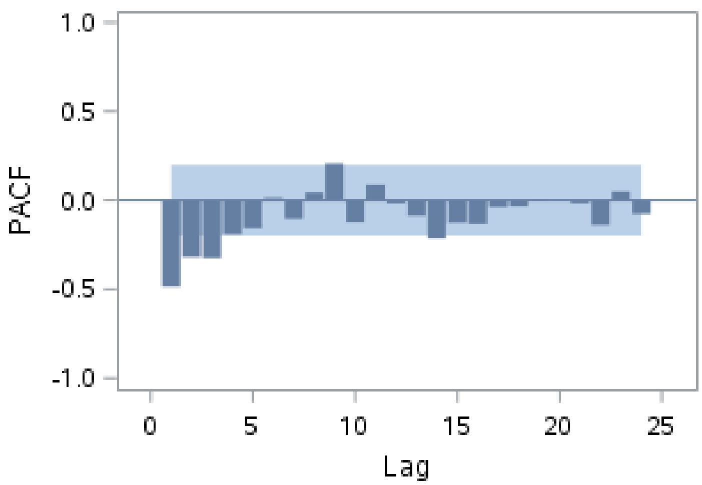


Both graphs(ACF and PACF) appear to be exponentially decaying to zero with the PACF graph alternating between positive and negative numbers. I classify this series as being generated by an **ARMA(1,1)** process.

PROBLEM 5)

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	26.76	6	0.0002	-0.486	-0.009	-0.062	0.092	-0.032	0.085
12	43.88	12	<.0001	-0.146	0.136	0.035	-0.230	0.215	-0.099
18	51.37	18	<.0001	0.004	-0.069	0.111	-0.073	0.113	-0.160
24	58.71	24	<.0001	0.080	-0.002	-0.032	-0.040	0.165	-0.137

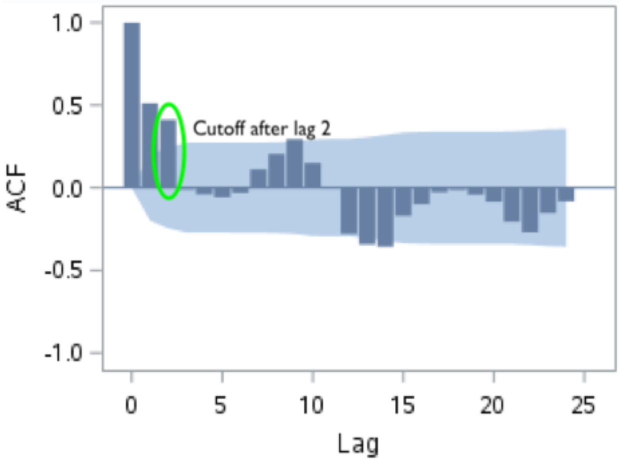


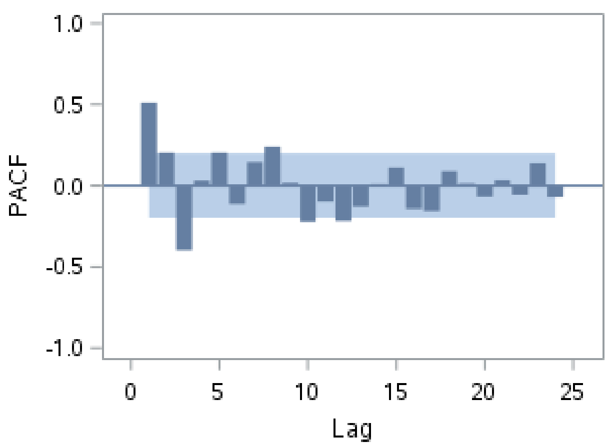


This series was generated by a **MA(1)** because the ACF has a cutoff after lag 1 and the PACF seems to decay exponentially to zero.

PROBLEM 6)

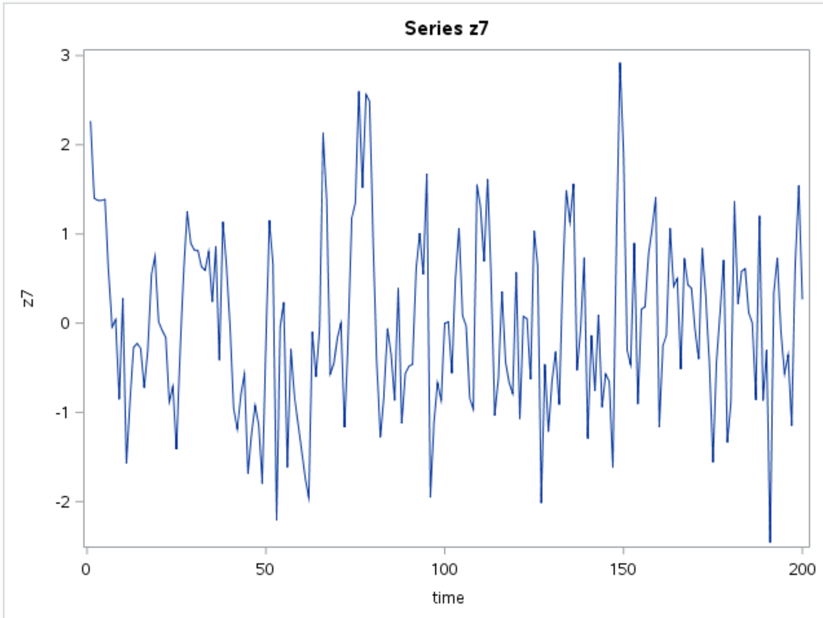
Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	45.68	6	<.0001	0.512	0.412	-0.010	-0.039	-0.057	-0.030
12	73.20	12	<.0001	0.112	0.204	0.294	0.151	-0.006	-0.279
18	107.24	18	<.0001	-0.343	-0.358	-0.170	-0.100	-0.026	-0.012
24	127.51	24	<.0001	-0.042	-0.085	-0.205	-0.270	-0.153	-0.083



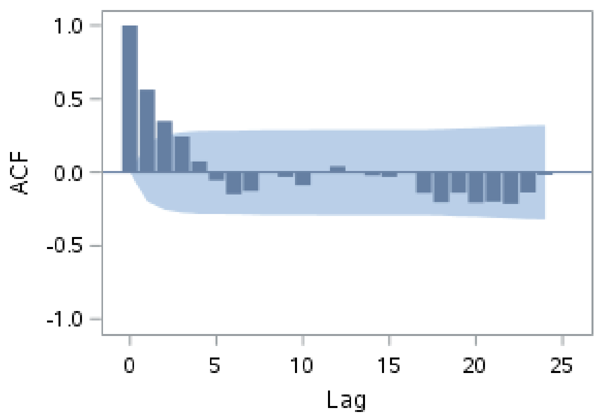


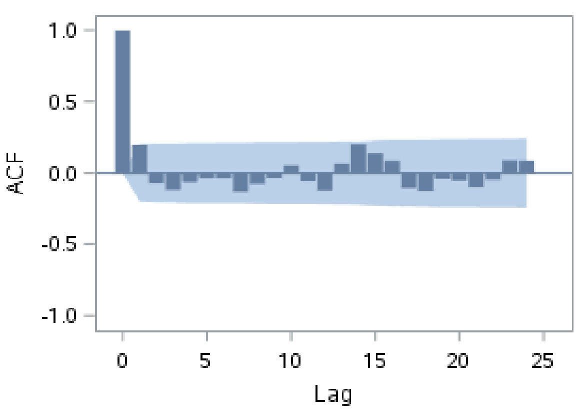
This series was generated by a **MA(2)** because the ACF has a cutoff at lag 2 and the PACF seems decay exponentially to zero alternating between positive and negative numbers

PROBLEM 7)



ACF for first half of the observations(1-100) ACF for second half of the observations(101-200)

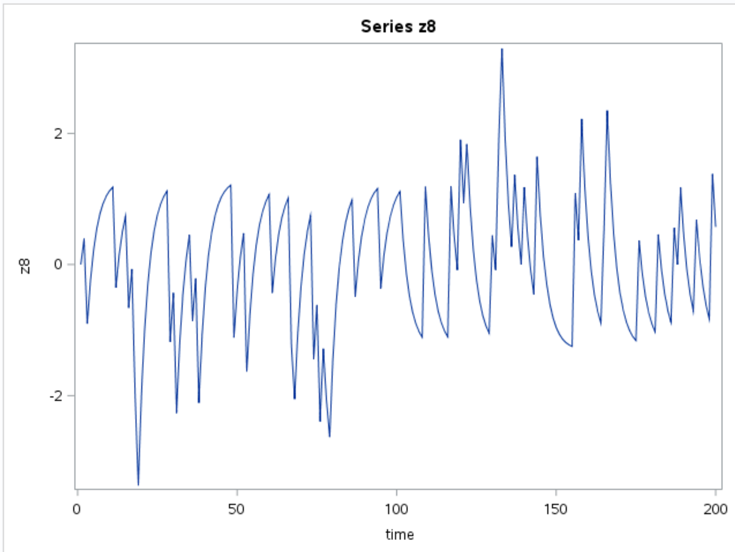




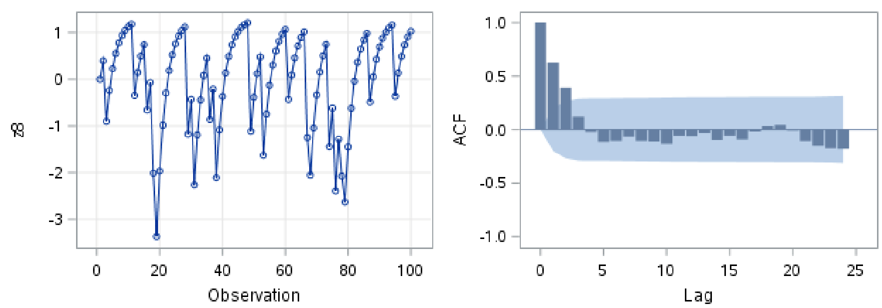
(d) Does not have a constant ACF.

The graph seems to indicate a constant mean and variance but we can observe that the ACF for the first half of the series(observations 1-100) is different from the ACF from the other half of the series(observations101-200 indicating the ACF is not constant throughout this series

PROBLEM 8)

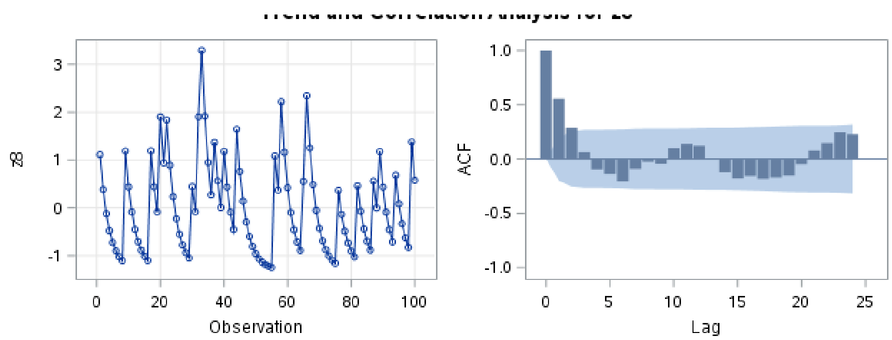


First half of times series z8(observations 1-100):



Name of Variable = z8	
Mean of Working Series	-0.03689
Standard Deviation	1.031988
Number of Observations	100

Second half of times series z8(observations 101-200):

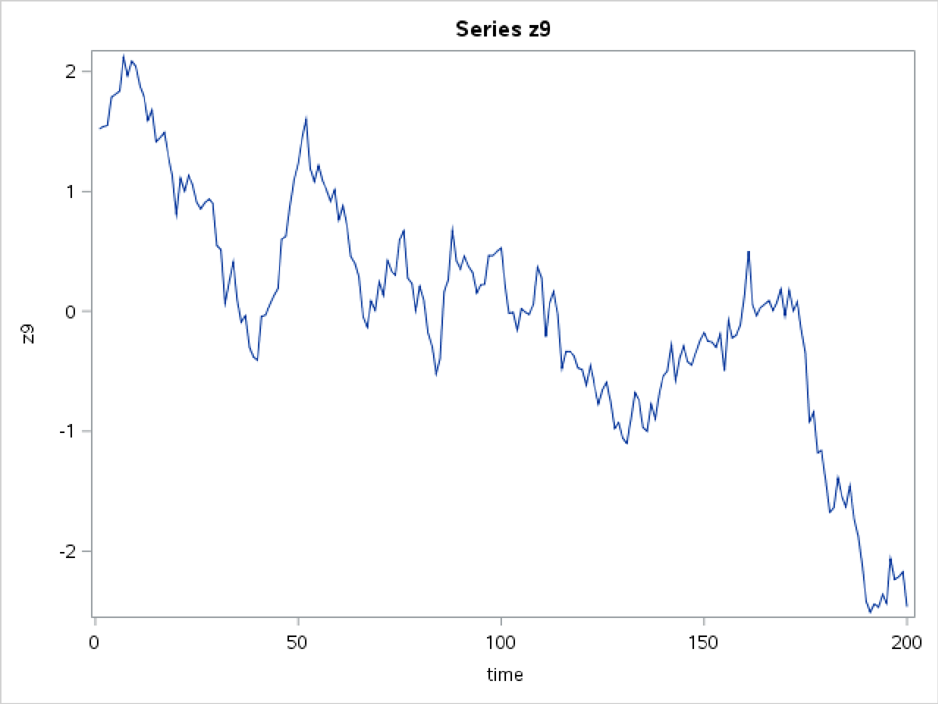


Name of Variable = z8	
Mean of Working Series	0.036886
Standard Deviation	0.96035
Number of Observations	100

(e) Is weakly stationary, but not strictly stationary

We can observe that this graph seems to have a constant variance, mean and ACF in both halves of the series but the behavior seems different on both halves so it is weakly stationary.

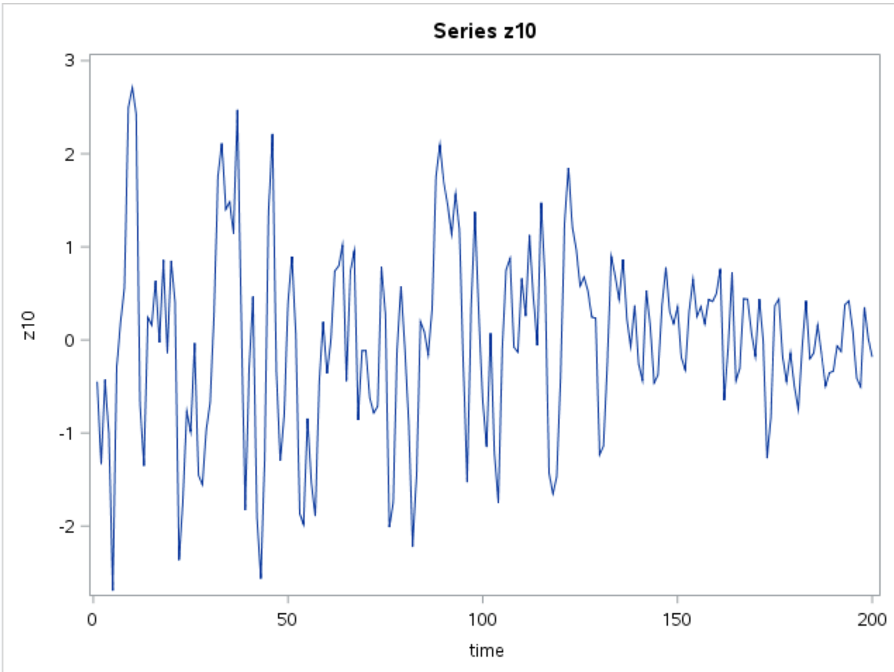
PROBLEM 9)



(b)Does not have a constant mean.

The graph does not seem to have a spread around 0 with values decreasing over time meaning the mean is not constant throughout the series.

PROBLEM 10)

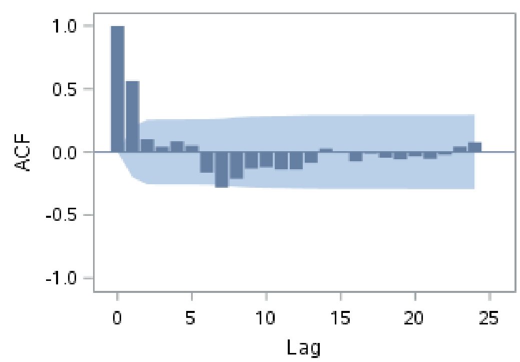


First half of series(observations 1 to 100):

Series z10, observations 1 to 100

The ARIMA Procedure

Name of Variable = z10	
Mean of Working Series	-0.05642
Standard Deviation	1.244017
Number of Observations	100

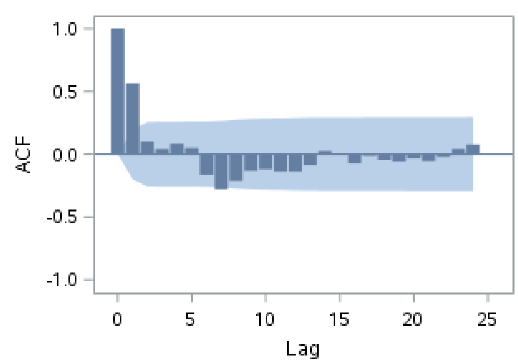


Second half of series(observations 101 to 200):

Series z10, observations 101 to 200

The ARIMA Procedure

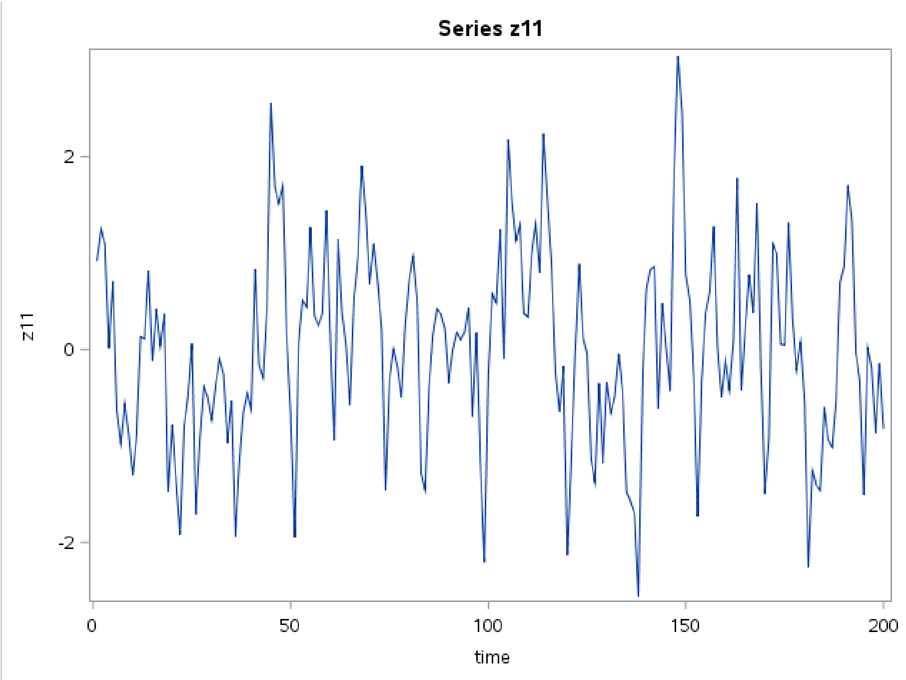
Name of Variable = z10	
Mean of Working Series	0.056416
Standard Deviation	0.660357
Number of Observations	100

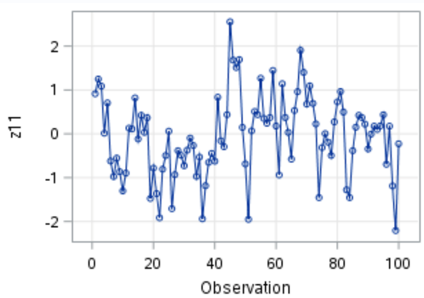


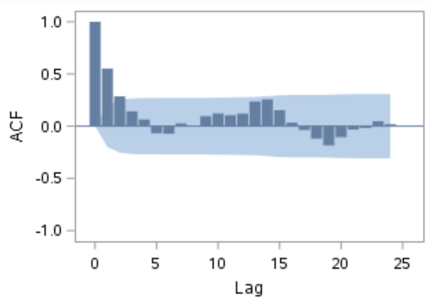
(c) Does not have a constant variance.

The mean and ACF seems to be constant and the time series graph display a spread around 0 throughout the entire series. However, the trend indicates that the variance gets smaller over time. Using the standard deviation from both halves, we can observe that the second half has a variance ½ times smaller than the first half which is in accordance with what the graph displays therefore the variance is not constant.

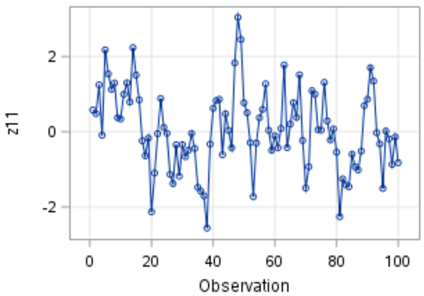
PROBLEM 11)

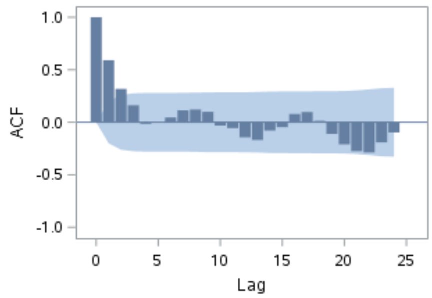






Name of Variable = z11	
Mean of Working Series	-0.05025
Standard Deviation	0.911515
Number of Observations	100





Name of Variable = z11	
Mean of Working Series	0.050241
Standard Deviation	1.074281
Number of Observations	100

(a) Stationary

The mean, ACF and variance are constant and the behavior seems to be the same on throughout all the so this series is stationary.

SAS Code

```
/* Problems 1-6 */
filename what "/home/eff100/my_courses/huffer/hw2p1_data.txt";
data hw2p1;
infile what;
input z1-z6;
run;

/* The following code will produce the usual items
  used to "identify" an ARMA model for a time series.
  This will be done for each of the series z1 to z6.  */

proc arima data=hw2p1;
identify var=z1; /* Problem 1 */
identify var=z2; /* Problem 2 */
identify var=z3; /* Problem 3 */
identify var=z4; /* Problem 4 */
identify var=z5; /* Problem 5 */
identify var=z6; /* Problem 6 */
run;

/* Problems 7-11 */
filename what "/home/eff100/my_courses/huffer/hw2p2_data.txt";

data look;
infile what;
time=_n_;
input z7-z11;
run;
/* Problem 7 */
/* Creating time series plots for z7 to z11. */
title "Series z7";
proc sgplot data=look;
series x=time y=z7;
run;

/* Splitting series z7 into half */
title "Series z7, observations 1 to 100";
proc arima data=look(firstobs=1 obs=100);
identify var=z7;
run;

title "Series z7, observations 101 to 200";
proc arima data=look(firstobs=101 obs=200);
identify var=z7;
run;

/* Problem 8 */
title "Series z8";
proc sgplot data=look;
series x=time y=z8;
run;

/* Splitting series z8 into half */
title "Series z8, observations 1 to 100";
proc arima data=look(firstobs=1 obs=100);
identify var=z8;
run;
```

```
title "Series z8, observations 101 to 200";
proc arima data=look(firstobs=101 obs=200);
identify var=z8;
run;

/* Problem 9 */
title "Series z9";
proc sgplot data=look;
series x=time y=z9;
run;

/* Problem 10 */
title "Series z10";
proc sgplot data=look;
series x=time y=z10;
run;

/* Splitting series z10 into half */
title "Series z10, observations 1 to 100";
proc arima data=look(firstobs=1 obs=100);
identify var=z10;
run;

title "Series z10, observations 101 to 200";
proc arima data=look(firstobs=101 obs=200);
identify var=z10;
run;

/* Problem 11 */
title "Series z11";
proc sgplot data=look;
series x=time y=z11;
run;

/* Splitting series z11 into half */
title "Series z11, observations 1 to 100";
proc arima data=look(firstobs=1 obs=100);
identify var=z11;
run;

title "Series z11, observations 101 to 200";
proc arima data=look(firstobs=101 obs=200);
identify var=z11;
run;
```