Homework 2

STA 5104

Allen Rahrooh

12 November 2019

Question 1

Part A)

```
DATA GDP1;
INFILE "/folders/myfolders/Homework 2/GDP.csv" DSD DLM = "," TRUNCOVER;
INPUT Country_Code $ Country_Name :$34. GDP 10.;
RUN;

PROC PRINT DATA = GDP1 NOOBS;
TITLE 'GDP BY COUNTRY';
RUN;
```

GDP BY COUNTRY

Country_Code	Country_Name	GDP
USA	United States	17946996
CHN	China	10866444
JPN	Japan	4123258
DEU	Germany	3355772
GBR	United Kingdom	2848755
FRA	France	2421682
IND	India	2073543
ITA	Italy	1814763
BRA	Brazil	1774725
CAN	Canada	1550537
KOR	Korea, Rep.	1377873
AUS	Australia	1339539
RUS	Russian Federation	1326015
ESP	Spain	1199057
MEX	Mexico	1144331
IDN	Indonesia	861934
NLD	Netherlands	752547
TUR	Turkey	718221
CHE	Switzerland	664738
SAU	Saudi Arabia	646002
ARG	Argentina	583169
SWE	Sweden	492618
NGA	Nigeria	481066
POL	Poland	474783
BEL	Belgium	454039
IRN	Iran, Islamic Rep.	425326
THA	Thailand	395282

Figure 1 GDP Data

```
DATA GEP;
INFILE "/folders/myfolders/Homework 2/GEP_DATA.csv" DSD DLM = "," TRUNCOVER;
INPUT Country_Name :$34. Country_Code $ y_2001 y_2002 y_2003 y_2004 y_2005 y_2006
y_2007 y_2008 y_2009 y_2010 y_2011 y_2012 y_2013 y_2014 y_2015 y_2016 y_2017 y_2018;

PROC PRINT DATA = GEP NOOBS;
TITLE 'GEP BY COUNTRY';
RUN;
```

GEP BY COUNTRY

Country_Name	Country_Code	y_2001	y_2002	y_2003	y_2004	y_2005	y_2006	y_2007	y_2008	y_2009	y_2010	y_2011	y_2012	y_2013	y_2014	y_2015	y_2016	y_2017	y_2018
Afghanistan	AFG	20.0641	16.7111	14.3184	9.4388	14.5157	11.1870	11.1320	3.4000	20.4000	8.4000	6.1000	14.400	2.0000	1.3000	1.9000	3.1000	3.9000	5.00000
Albania	ALB	7.0000	2.9000	5.7000	5.9000	5.4999	5.0000	5.9001	7.5363	3.3542	3.7069	2.5454	1.624	1.4174	2.0200	2.7000	3.4000	3.5000	3.50000
Algeria	DZA	4.6125	5.6000	7.2000	4.3000	5.9850	1.7000	3.4000	2.4000	1.6000	3.6002	2.9000	3.400	2.8000	3.8000	2.8000	3.9000	4.0000	3.80000
Angola	AGO	4.2210	13.8218	5.2476	10.8795	18.2615	20.7351	22.5931	13.8171	2.4129	3.4077	3.9186	5.155	6.8001	3.9013	3.0378	3.3060	3.8168	3.83121
Antigua and Barbuda	ATG	-3.1942	2.9243	5.9136	5.2864	6.0832	13.3764	9.4989	0.0711	-12.0360	-7.1430	-1.7934	4.020	-0.0714	3.2215	2.0111	2.4224	2.7349	2.73491
Argentina	ARG	-4.4088	-10.8945	8.8370	9.0296	9.2263	8.3752	7.9656	3.0749	0.0500	9.4516	8.3865	0.802	2.8854	0.4536	1.6598	0.6693	1.9268	2.99641
Armenia	ARM	9.5566	13.1863	14.0408	10.4678	13.8657	13.1980	13.7492	6.9000	-14.1500	2.2000	4.7000	7.200	3.3000	3.5000	2.5000	2.2000	2.8000	3.00000
Australia	AUS	2.5745	3.9966	3.0208	4.0358	3.2142	2.6546	4.5199	2.6711	1.5747	2.2501	2.7217	3.600	2.4400	2.5000	2.1870	2.5000	2.7500	3.00000
Austria	AUT	1.3505	1.6559	0.7561	2.7057	2.1407	3.3508	3.6215	1.5473	-3.7991	1.8801	3.0714	0.884	0.2280	0.3007	0.7000	1.5000	1.6000	1.50000
Azerbaijan	AZE	9.9000	10.6000	11.2000	10.2000	26.4000	34.5000	25.0490	10.7724	9.4107	4.8543	0.0659	2.200	5.7967	2.8209	1.9664	0.7726	1.2061	2.66506
Bahamas	BHS	2.6256	2.7046	-1.2647	0.8829	3.3953	2.5169	1.4465	-2.3239	-4.1753	1.5388	0.6129	2.217	0.0220	1.0228	2.1069	2.0000	1.7000	1.70000
Bahrain	BHR	2.4909	3.3486	6.2964	6.9810	6.7690	6.4670	8.2940	6.2418	2.5405	4.3368	2.1003	3.589	5.4086	4.4859	2.5000	2.7000	2.7000	2.80000
Bangladesh	BGD	3.8332	4.7396	5.2395	6.5359	6.6719	7.0586	6.0138	5.0451	5.5718	6.4644	6.5215	6.014	6.1162	6.4546	6.5000	6.7000	6.8000	6.80000
Barbados	BRB	-2.3673	0.7911	2.1713	1.4068	3.9650	5.6686	1.7643	0.3960	-4.0323	0.2558	0.7562	0.280	-0.0361	0.2075	0.4929	1.3769	2.0319	2.03192
Belarus	BLR	4.7253	5.0453	7.0432	11.4497	9.4000	10.0000	8.6000	10.2000	0.2000	7.7408	5.5437	1.731	1.0736	1.5878	-3.4656	-0.5000	0.9643	0.96430
Belgium	BEL	0.9125	1.5593	0.8889	3.4334	1.9036	2.6308	2.9915	0.9514	-2.6162	2.5006	1.6275	0.090	0.2852	1.0588	1.3023	1.5231	1.6130	1.51321
Belize	BLZ	5.0157	5.1186	9.3292	4.6481	2.5780	4.5816	1.1056	3.2291	0.7133	3.3239	2.1040	3.825	1.5257	3.5730	3.0092	2.4931	2.6001	2.82529
Benin	BEN	6.2484	4.4211	3.8825	3.1199	2.8652	3.7522	4.6264	5.0146	2.6632	2.6115	3.2637	5.396	5.6401	5.3691	5.7072	5.3084	5.0733	5.08617
Bhutan	BTN	6.7400	9.5160	9.1180	6.7480	6.5270	6.9810	12.5710	10.7520	5.7220	9.3270	10.0690	3.600	3.9000	6.3000	6.8000	7.2000	5.6000	6.00000
Bolivia	BOL	1.6838	2.4856	2.7113	4.1733	4.4214	4.7970	4.5644	6.1485	3.3570	4.1267	5.1739	5.108	6.8425	5.4871	3.9765	3.5144	3.4075	3.39827
Bosnia and Herzegovina	BIH	4.4000	5.3000	4.0000	6.1000	5.0000	6.2000	6.8380	5.4200	-2.9100	0.8000	1.0000	-1.200	2.5000	0.8000	1.9000	2.3000	3.1000	3.50000
Botswana	BWA	0.2506	6.0695	4.6259	2.7058	4.5566	7.9595	8.6823	6.2453	-7.6522	8.5633	6.0485	4.831	9.3224	4.4211	2.9937	4.0198	4.1622	4.16312
Brazil	BRA	1.2922	3.0422	1.2359	5.6476	3.7392	4.0315	6.0979	5.0707	-0.0659	7.5241	3.9265	1.886	3.0022	0.1257	-3.7425	-2.4614	1.3725	1.53200
Bulgaria	BGR	4.2481	6.0180	5.0775	6.5556	9.0334	6.7533	7.6752	5.6471	-4.2197	0.0545	1.5840	0.237	1.2818	1.5488	2.9372	2.1684	2.6851	2.68423
Burkina Faso	BFA	6.6134	4.3530	7.8024	4.4785	8.6731	6.2532	5.6550	7.2945	2.9620	8.4463	6.5218	6.453	6.6694	4.0498	4.3992	5.9928	6.9810	6.98544
Burundi	BDI	2.0558	4.4465	-1.2237	4.8337	0.9000	5.3847	4.7858	5.0481	3.4684	3.7859	4.1916	4.019	4.5941	4.6609	-2.2996	3.4829	4.7783	4.76989
Cambodia	KHM	8.0386	6.6879	8.5059	10.3405	13.2501	10.7711	10.2126	6.6916	0.0867	5.9631	7.0696	7.313	7.3567	7.0317	6.9318	6.9272	6.8246	6.82463
Cameroon	CMR	4.5143	4.0090	4.0310	3.7019	2.2969	3.2240	3.2557	2.8840	1.9319	3.2686	4.1406	4.589	5.5617	5.8863	6.2913	6.4503	6.4793	6.36535
Canada	CAN	1.6885	2.8019	1.9253	3.1389	3.1631	2.6218	2.0083	1.1754	-2.7114	3.3742	2.9601	1.923	2.0035	2.4393	1.2000	1.9000	2.3000	2.40000
Cape Verde	CPV	2.2317	5.2509	4.1763	10.1971	6.9235	7.9838	15.1707	6.6505	-1.2704	1.4669	3.9688	1.082	1.0452	1.8140	2.9228	3.4621	4.1181	4.14418
Chad	TCD	11.6581	8.4912	14.7217	33.6294	17.3325	0.6483	3.2715	3.0527	4.2177	13.5501	0.0829	8.883	5.7000	7.3000	4.1275	4.9179	6.0888	6.52775
Chile	CHL	3.3482	2.1669	3.9567	6.7970	7.1327	5.6595	5.2074	3.2942	-1.0392	5.8067	5.7991	5.513	4.1839	1.8759	2.1420	2.4212	2.8569	3.08991
China	CHN	8.2984	9.0909	10.0200	10.0756	11.3524	12.6882	14.1950	9.6234	9.2335	10.6317	9.4845	7.750	7.6838	7.3164	6.8945	6.7190	6.4544	6.47584
Colombia	COL	1.6779	2.5040	3.9183	5.3330	4.7066	6.6975	6.9006	3.5468	1.6515	3.9718	6.5895	4.044	4.9363	4.5525	3.1298	2.9609	3.3399	3.54017
Comoros	COM	11.8482	2.3248	2.1041	1.9196	2.8480	2.6476	0.7997	0.4000	1.9500	2.2000	2.6000	3.000	3.5000	3.0037	2.3139	2.4610	3.0978	3.09535
Congo, Dem. Rep.	ZAR	-2.1002	2.9478	5.5778	6.7384	6.1352	5.3210	6.2595	6.2259	2.8551	7.0789	6.8646	7.158	8.5035	9.0466	7.9697	8.5694	9.0026	9.00264
Congo, Rep.	COG	3.8026	4.5819	0.8133	3.4766	7.7558	6.2360	-1.5822	5.5723	7.4689	8.7517	3.4206	3.800	3.4407	6.5462	1.2927	3.4642	5.5883	5.59230
Costa Rica	CRI	1.0764	2.9022	6.4045	4.2595	5.8864	8.7796	7.9353	2.7316	-1.0157	4.9543	4.5177	5.168	3.4373	3.5024	2.8213	3.9744	4.1682	4.39017

Figure 2 GEP Data

Part B)

```
PROC SGPLOT data=WORK.GDP1;
        title height=14pt "GDP HISTOGRAM";
        histogram GDP / nbins=50 fillattrs=(color=CX804040) filltype=gradient
                dataskin=gloss;
        density GDP / type=Kernel;
        xaxis grid;
        yaxis grid;
        keylegend "DENSITY" / location=inside position=topright across=1;
run;
```

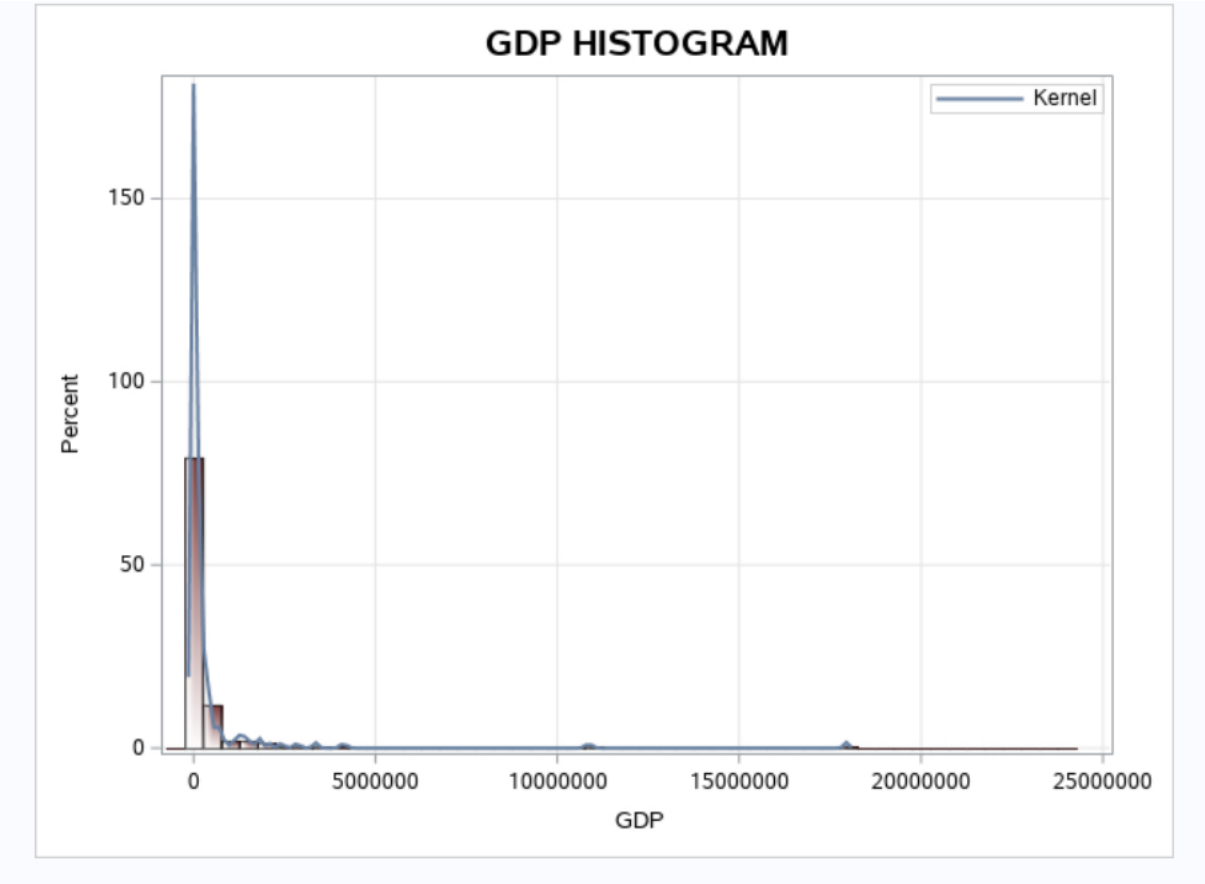

Figure 3 GDP Histogram

Part C)

```
proc sort data=GDP1 out=GDP_Descending;
by descending GDP;
run;

proc sql;
create table TOPTEN as
select *
from GDP_Descending (obs =10);
quit;


PROC sgplot data=WORK.TOPTEN;
	title height=14pt "TOP TEN GDP COUNTRIES";
	vbar Country_Code / response=GDP fillattrs=(color=CX3bd02d) datalabel
		fillType=gradient stat=percent dataskin=gloss;
	yaxis grid;
run;
```

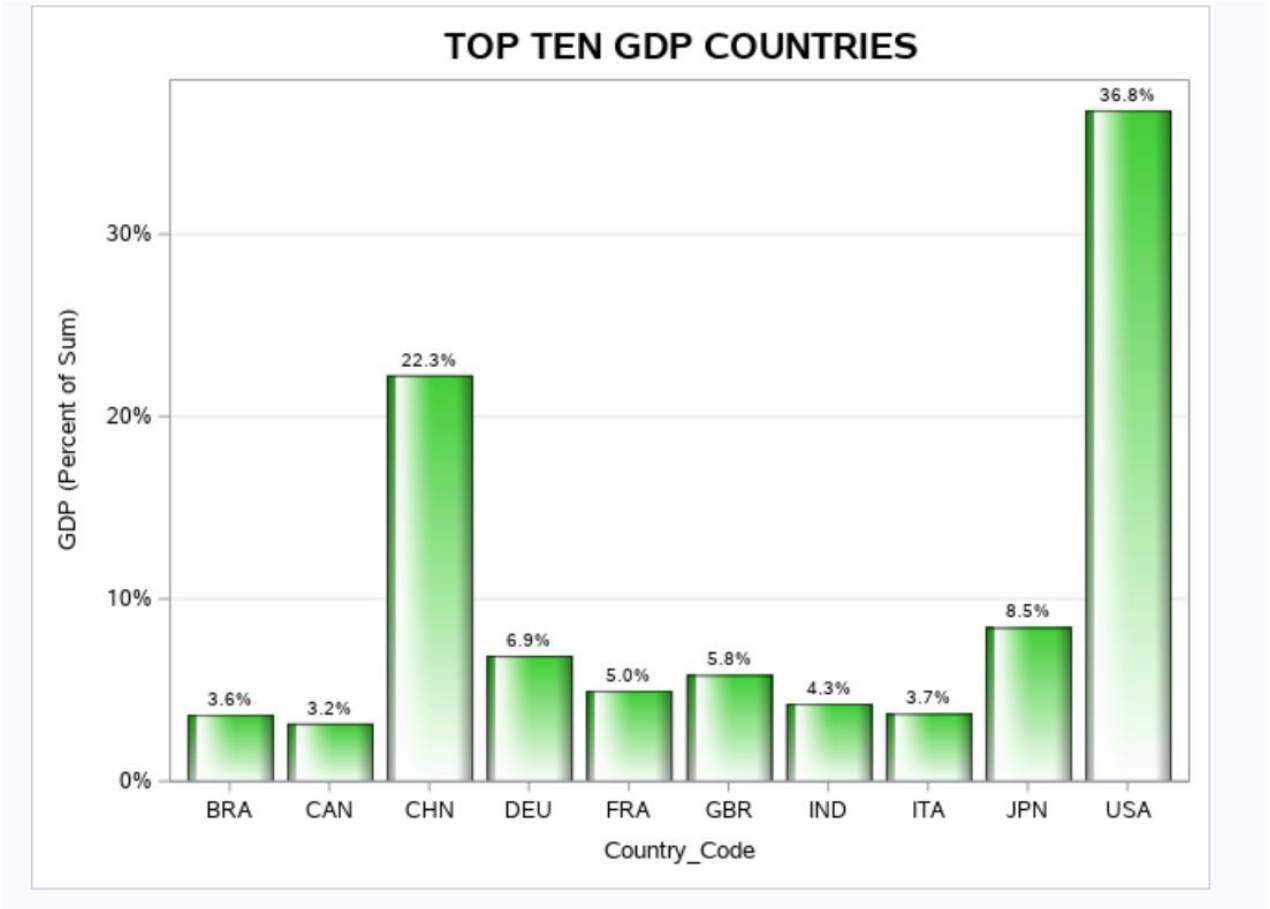

Figure 4 Vertical Bar Graph of Top Ten Countries

Part D)

```
PROC SQL;
CREATE TABLE GEP1 AS
SELECT *
        FROM GEP
        WHERE Country_Code
        IN ('USA', 'GRC', 'CHN', 'GBR', 'ARG');
RUN;

PROC PRINT DATA = GEP1 NOOBS;
TITLE 'FIVE SELECTED COUNTRIES GEP';
```

FIVE SELECTED COUNTRIES GEP

Country_Name	Country_Code	y_2001	y_2002	y_2003	y_2004	y_2005	y_2006	y_2007	y_2008	y_2009	y_2010	y_2011	y_2012	y_2013	y_2014	y_2015	y_2016	y_2017	y_2018
Argentina	ARG	-4.40884	-10.8945	8.8370	9.0296	9.2263	8.3752	7.9656	3.07494	0.05002	9.4516	8.38645	0.80176	2.88535	0.45360	1.65975	0.66926	1.92680	2.99641
China	CHN	8.29837	9.0909	10.0200	10.0756	11.3524	12.6882	14.1950	9.62338	9.23355	10.6317	9.48451	7.75029	7.68382	7.31644	6.89451	6.71902	6.45437	6.47584
Greece	GRC	3.60954	3.1424	6.5379	4.8841	1.1330	5.7449	3.3822	-0.43795	-4.36041	-5.3369	-8.86572	-6.62053	-3.98094	0.69028	-2.10000	-0.90000	0.80000	1.60000
United Kingdom	GBR	2.75796	2.4940	3.3367	2.4884	2.9964	2.6618	2.5861	-0.46688	-4.19194	1.5402	1.97240	1.17906	2.15990	2.94019	2.41321	2.40000	2.20110	2.10000
United States	USA	0.97615	1.7875	2.8058	3.7857	3.3463	2.6654	1.7791	-0.29112	-2.77604	2.5321	1.60107	2.22428	1.48945	2.42758	2.49080	2.70231	2.40213	2.20540

Figure 5 GEP of Five Selected Countries

Part E)

```
PROC TRANSPOSE DATA = GEP1 OUT = FIVECOUNTRY;
id Country_Code;
run;

DATA fivecountry2 (RENAME = (_NAME_ = YEAR));
SET FIVECOUNTRY;
RUN;

PROC SGPLOT data=WORK.FIVECOUNTRY2;
       title height=14pt "SCATTER PLOT OF USA GEP";
       scatter x=YEAR y=USA /;
       xaxis grid label="YEARS";
       yaxis grid;
run;
```

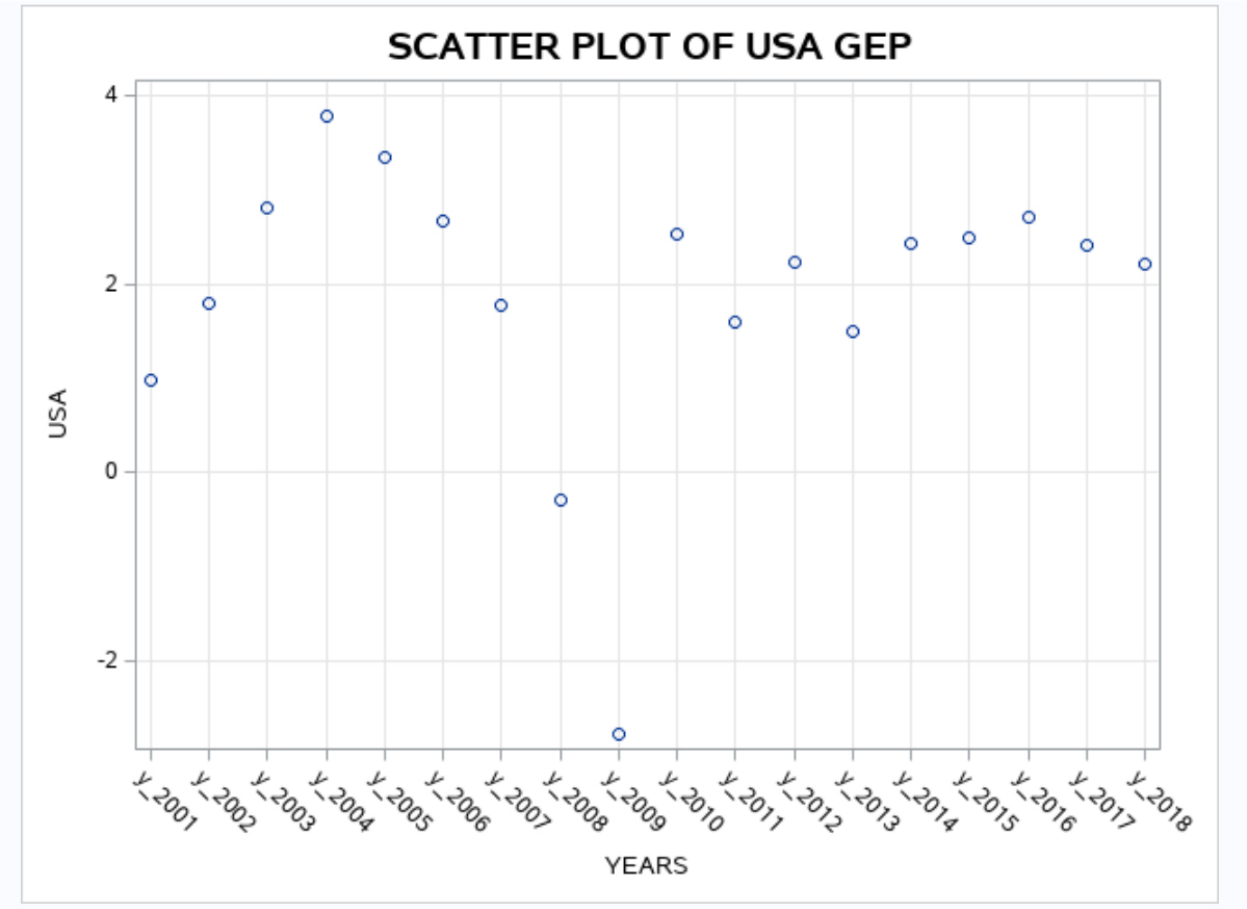

Figure 6 Scatter Plot of USA GEP

```
PROC SGPLOT data=WORK.FIVECOUNTRY2;
       title height=14pt "SCATTER PLOT OF GRC GEP";
       scatter x=YEAR y=GRC /;
       xaxis grid label="YEARS";
       yaxis grid;
run;
```

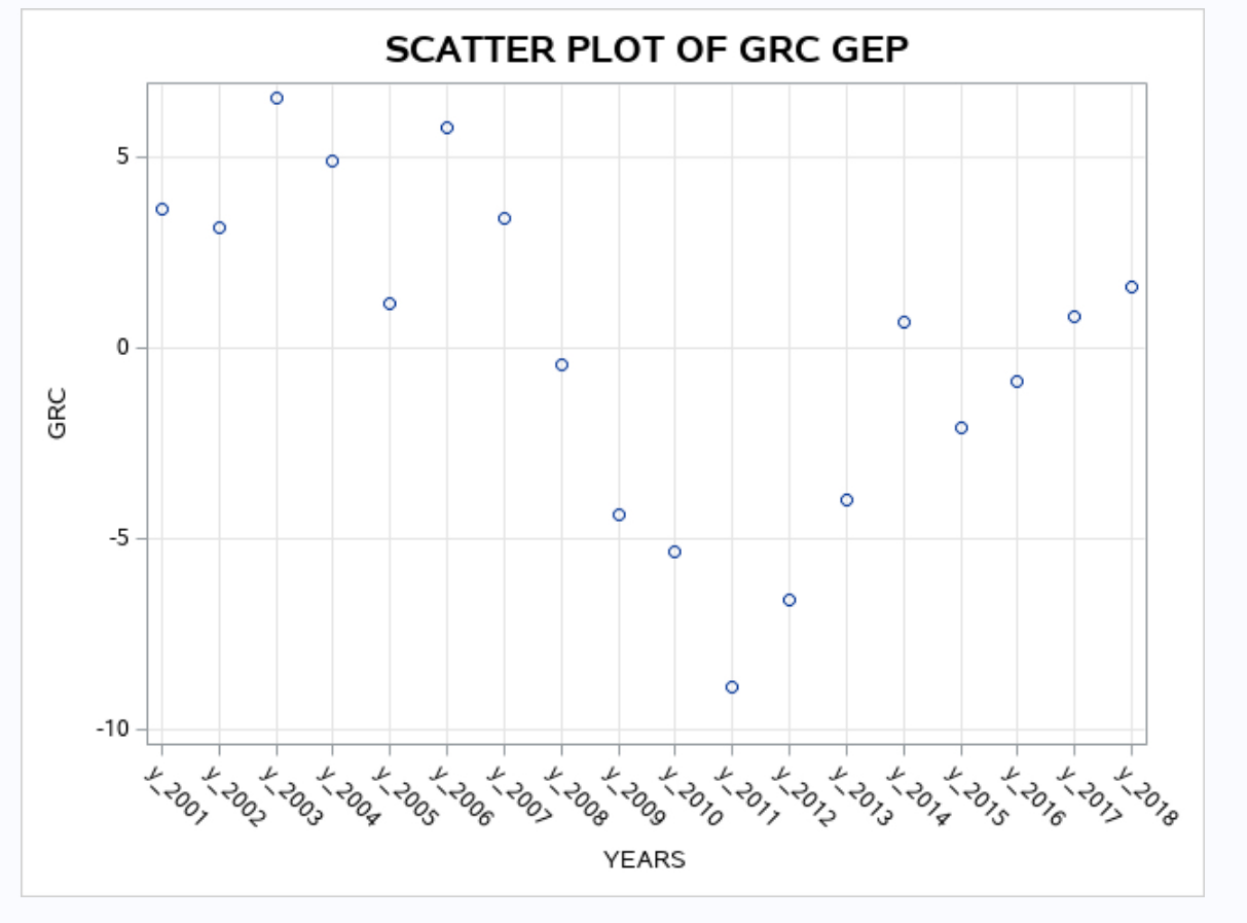

Figure 7 Scatter Plot of Greece GEP

```
PROC SGPLOT data=WORK.FIVECOUNTRY2;
       title height=14pt "SCATTER PLOT OF CHN GEP";
       scatter x=YEAR y=CHN /;
       xaxis grid label="YEARS";
       yaxis grid;
run;
```

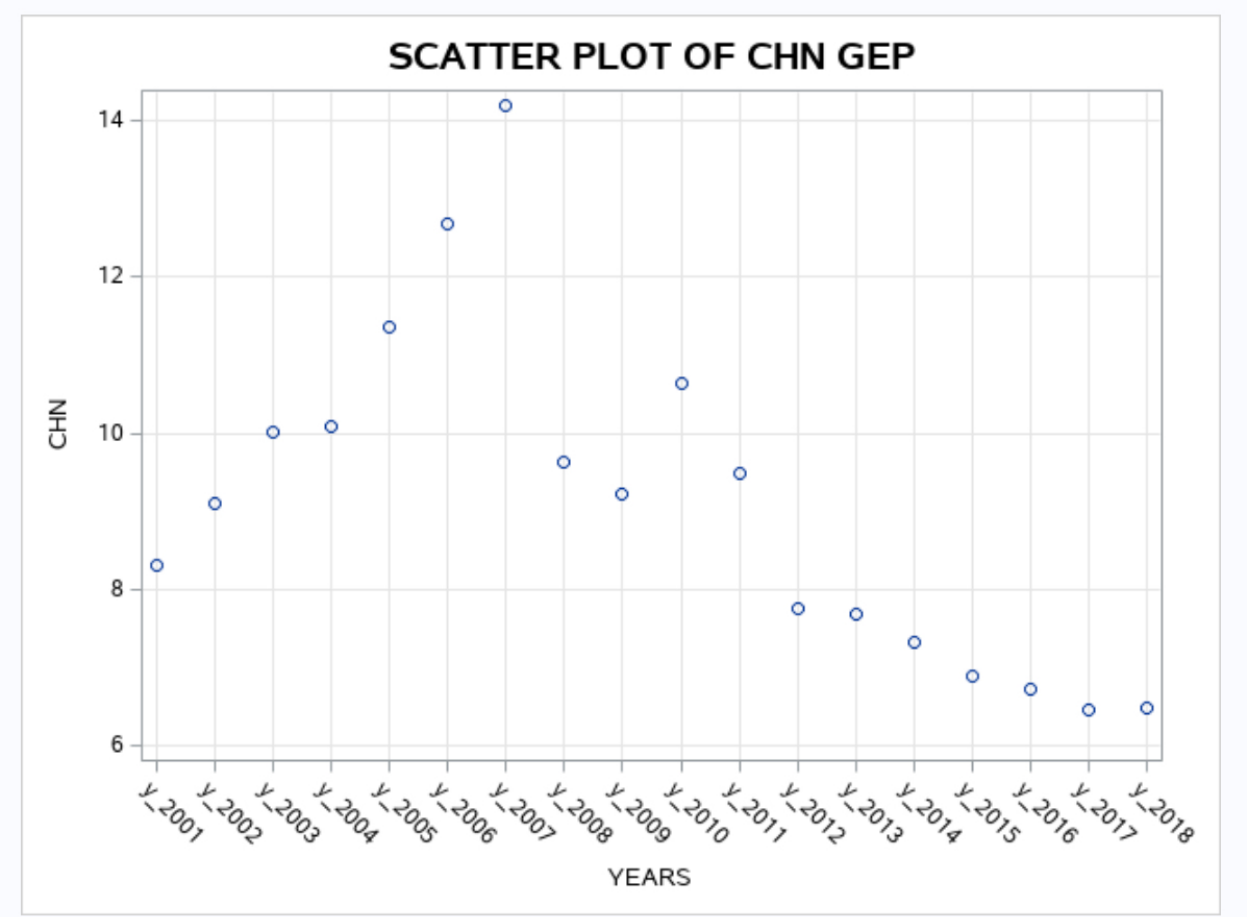

Figure 8 Scatter Plot of China GEP

```
PROC SGPLOT data=WORK.FIVECOUNTRY2;
       title height=14pt "SCATTER PLOT OF GBR GEP";
       scatter x=YEAR y= GBR /;
       xaxis grid label="YEARS";
       yaxis grid;
run;
```

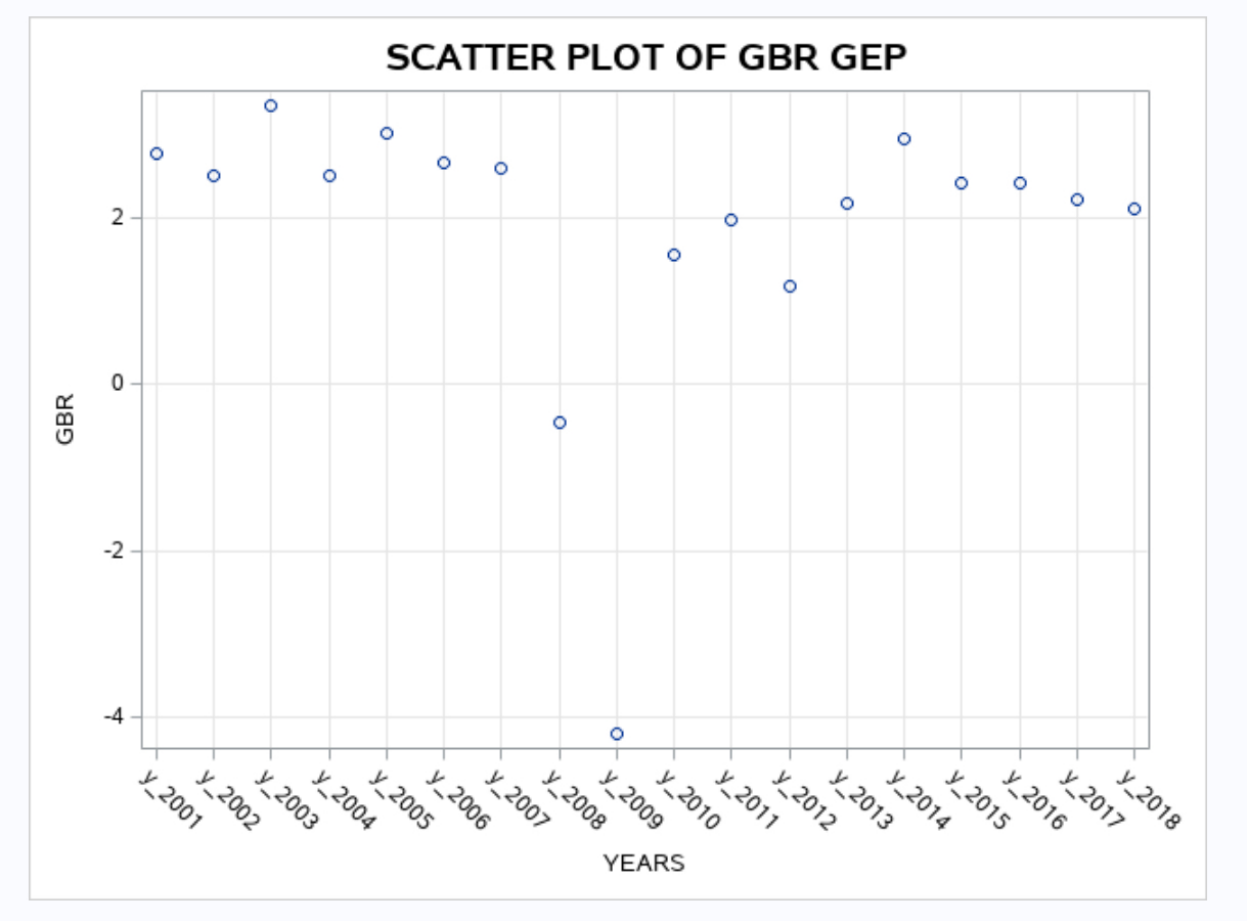

Figure 9 Scatter Plot of Great Britain GEP

```
PROC SGPLOT data=WORK.FIVECOUNTRY2;
       title height=14pt "SCATTER PLOT OF ARG GEP";
       scatter x=YEAR y=ARG /;
       xaxis grid label="YEARS";
       yaxis grid;
run;
```

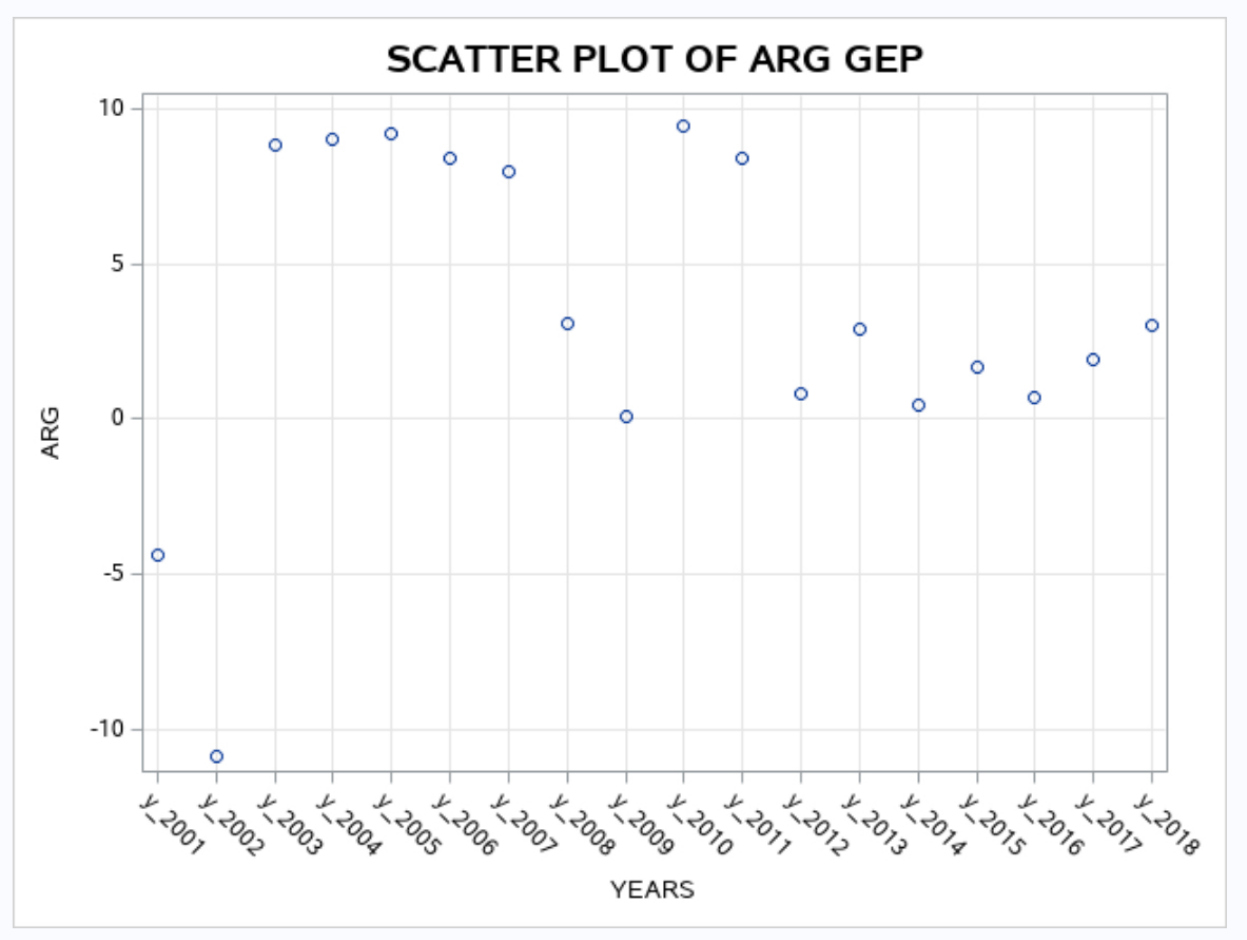

Figure 10 Scatter Plot of Argentina GEP

Part F)

```
PROC SGPLOT DATA = fivecountry2;
SERIES X = YEAR  Y = ARG/MARKERS;
SERIES X = YEAR  Y = USA/MARKERS;
SERIES X = YEAR  Y = GRC/MARKERS;
SERIES X = YEAR  Y = GBR/MARKERS ;
SERIES X = YEAR  Y = CHN/MARKERS ;
XAXIS LABEL = "YEAR";
YAXIS LABEL = "GEP";
TITLE height=14pt "REGRESSION PLOT OF FIVE SELECTED COUNTRIES GEP";
RUN;
```

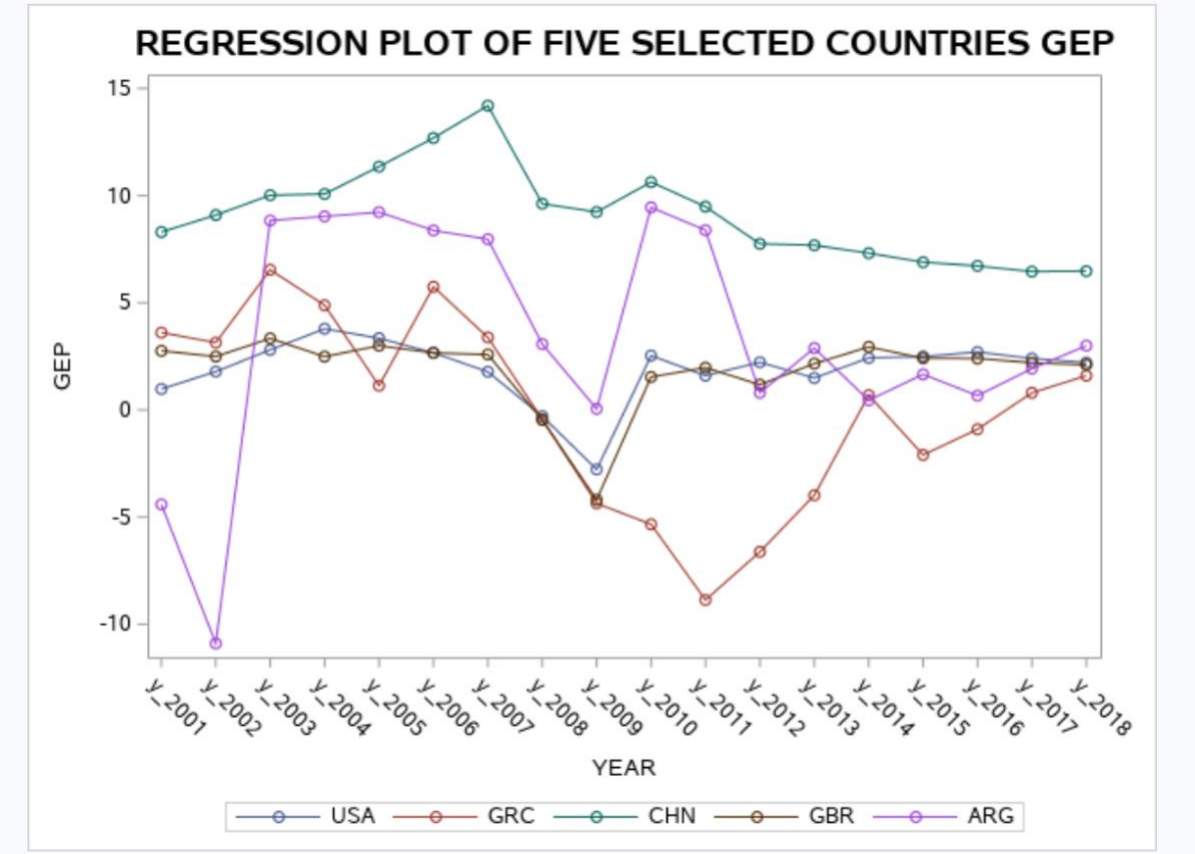

Figure 11 Regression Plot of Five Countries GEP

Part G)

```
PROC SQL;
CREATE TABLE WORK.QUERY
AS
SELECT GEP.Country_Code, GEP.y_2001, GEP.y_2002, GEP.y_2003, GEP.y_2004,
GEP.y_2005, GEP.y_2006, GEP.y_2007, GEP.y_2008, GEP.y_2009, GEP.y_2010,
GEP.y_2011, GEP.y_2012, GEP.y_2013, GEP.y_2014, GEP.y_2015, GEP.y_2016,
GEP.y_2017, GEP.y_2018, GDP1.GDP
FROM WORK.GEP GEP
LEFT JOIN WORK.GDP1 GDP1
ON
   ( GEP.Country_Code = GDP1.Country_Code) ;
QUIT;
```

Part H)

```
PROC SQL;
CREATE TABLE WORK.QUERY1
AS
SELECT MAX(GEP1.y_2001)
AS y_2001, MAX(GEP1.y_2002)
```

```
AS y_2002, MAX(GEP1.y_2003)
AS y_2003, MAX(GEP1.y_2004)
AS y_2004, MAX(GEP1.y_2005)
AS y_2005, MAX(GEP1.y_2006)
AS y_2006, MAX(GEP1.y_2007)
AS y_2007, MAX(GEP1.y_2008)
AS y_2008, MAX(GEP1.y_2009)
AS y_2009, MAX(GEP1.y_2010)
AS y_2010, MAX(GEP1.y_2011)
AS y_2011, MAX(GEP1.y_2012)
AS y_2012, MAX(GEP1.y_2013)
AS y_2013, MAX(GEP1.y_2014)
AS y_2014, MAX(GEP1.y_2015)
AS y_2015, MAX(GEP1.y_2016)
AS y_2016, MAX(GEP1.y_2017)
AS y_2017, MAX(GEP1.y_2018)
AS y_2018
FROM WORK.GEP1 GEP1;
QUIT;

PROC PRINT DATA = WORK.QUERY1;
TITLE "MAX PROJECTION PER YEAR";
```

MAX PROJECTION PER YEAR

Obs	y_2001	y_2002	y_2003	y_2004	y_2005	y_2006	y_2007	y_2008	y_2009	y_2010	y_2011	y_2012	y_2013	y_2014	y_2015	y_2016	y_2017	y_2018
1	8.29837	9.09091	10.0200	10.0756	11.3524	12.6882	14.1950	9.62338	9.23355	10.6317	9.48451	7.75029	7.68382	7.31644	6.89451	6.71902	6.45437	6.47584

Figure 12 Maximum Projection Per Year

```
PROC SGPLOT DATA = MAX;
SERIES X = _NAME_  Y= COL1/MARKERS;
XAXIS LABEL = "YEAR";
YAXIS LABEL = "GEP";
title height=14pt "REGRESSION PLOT OF MAXIMUM PROJECTIONS PER YEAR";
RUN;
```

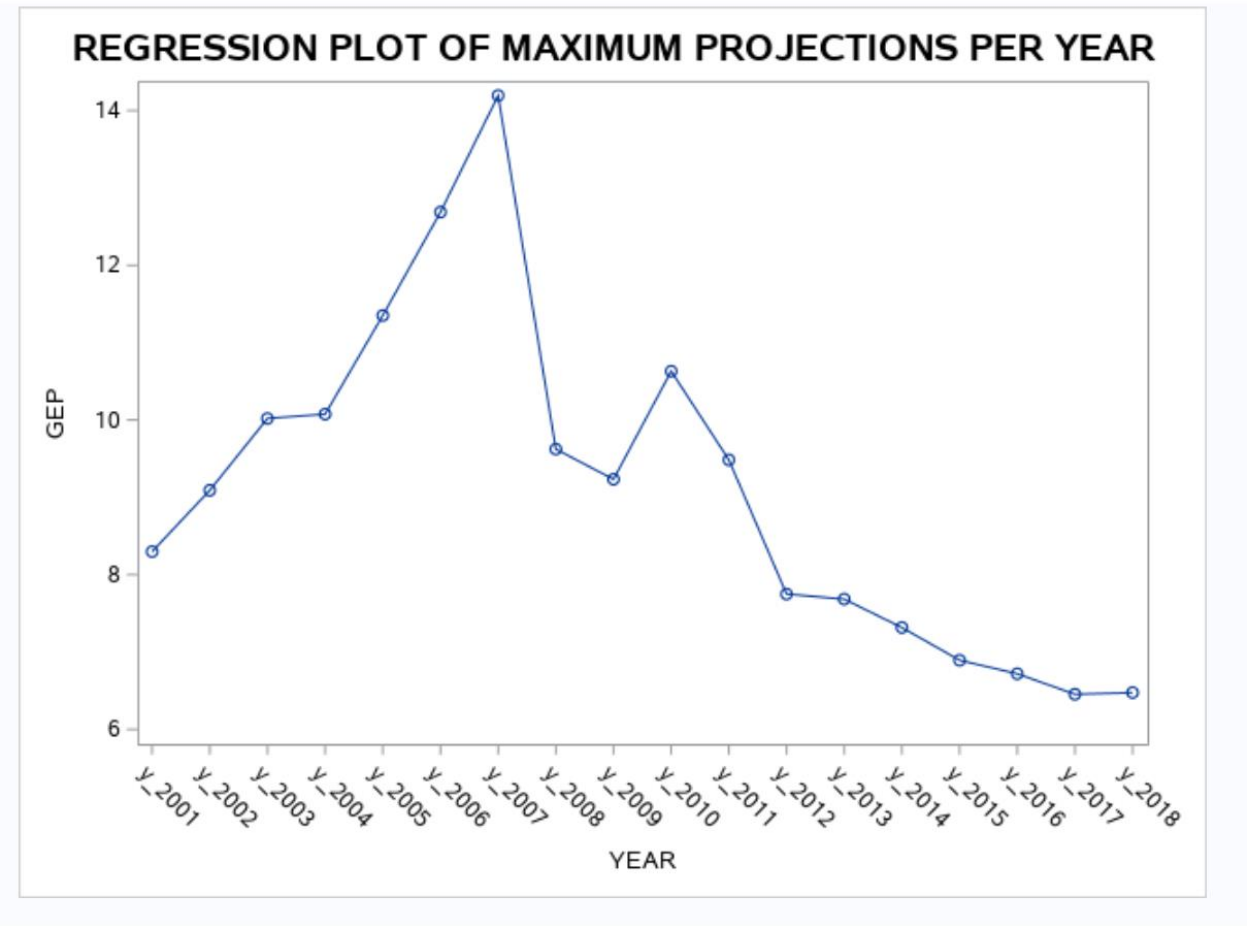

Figure 13 Regression Plot of Maximum Projections

Part I)

```
PROC CORR DATA = WORK.QUERY;
VAR GDP y_2007;
RUN;
```

The CORR Procedure

2 Variables:	GDP y_2007

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
GDP	172	418903	1675246	72051249	38.00000	17946996
y_2007	180	6.20465	4.07488	1117	-3.65333	25.04900

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	GDP	y_2007
GDP	1.00000 172	-0.04847 0.5290 171
y_2007	-0.04847 0.5290 171	1.00000 180

Figure 14 Correlation Between GDP and 2007

Part J)

What conclusions can you draw from these results?

According to the correlation coefficient -0.04847 GDP is negatively associated with GEP the projected growth for that year.

This means that, the more GDP, the less growth we would expect to see in 2007. Therefore, especially considering p = 0.5290 for this test, it is likely that countries with a high GDP had a lower rate of growth, or even saw their economies shrink for the year 2007.

Question 2

Part A)

```
FILENAME REFFILE '/folders/myfolders/Homework 2/GEPsupplementRecent.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=SUPPLEMENT;
        GETNAMES=YES;
RUN;
```

```
PROC SQL;
CREATE TABLE WORK.QUERY2
AS
SELECT SUPPLEMENT.Country_Name, SUPPLEMENT.Country_Code, SUPPLEMENT._2015__2015_,
GEP.y_2015, SUPPLEMENT._2016__2016_, GEP.y_2016, SUPPLEMENT._2017__2017_, GEP.y_2017,
SUPPLEMENT._2018__2018_, GEP.y_2018
FROM WORK.GEP GEP
INNER JOIN WORK.SUPPLEMENT SUPPLEMENT
ON
  ( GEP.Country_Name = SUPPLEMENT.Country_Name ) ;
QUIT;


DATA DIFFERENCE;

SET WORK.QUERY2;

diff_2015 = _2015__2015_ - y_2015;
diff_2016 = _2016__2016_ - y_2016;
diff_2017 = _2017__2017_ - y_2017;
diff_2018 = _2018__2018_ - y_2018;


KEEP Country_Code;
KEEP diff_2015;
KEEP diff_2016;
KEEP diff_2017;
KEEP diff_2018;
RUN;


PROC SGSCATTER DATA=DIFFERENCE;
    MATRIX diff_2015 diff_2016 diff_2017 diff_2018;
TITLE "SCATTER PLOT MATRIX ";
RUN;
```

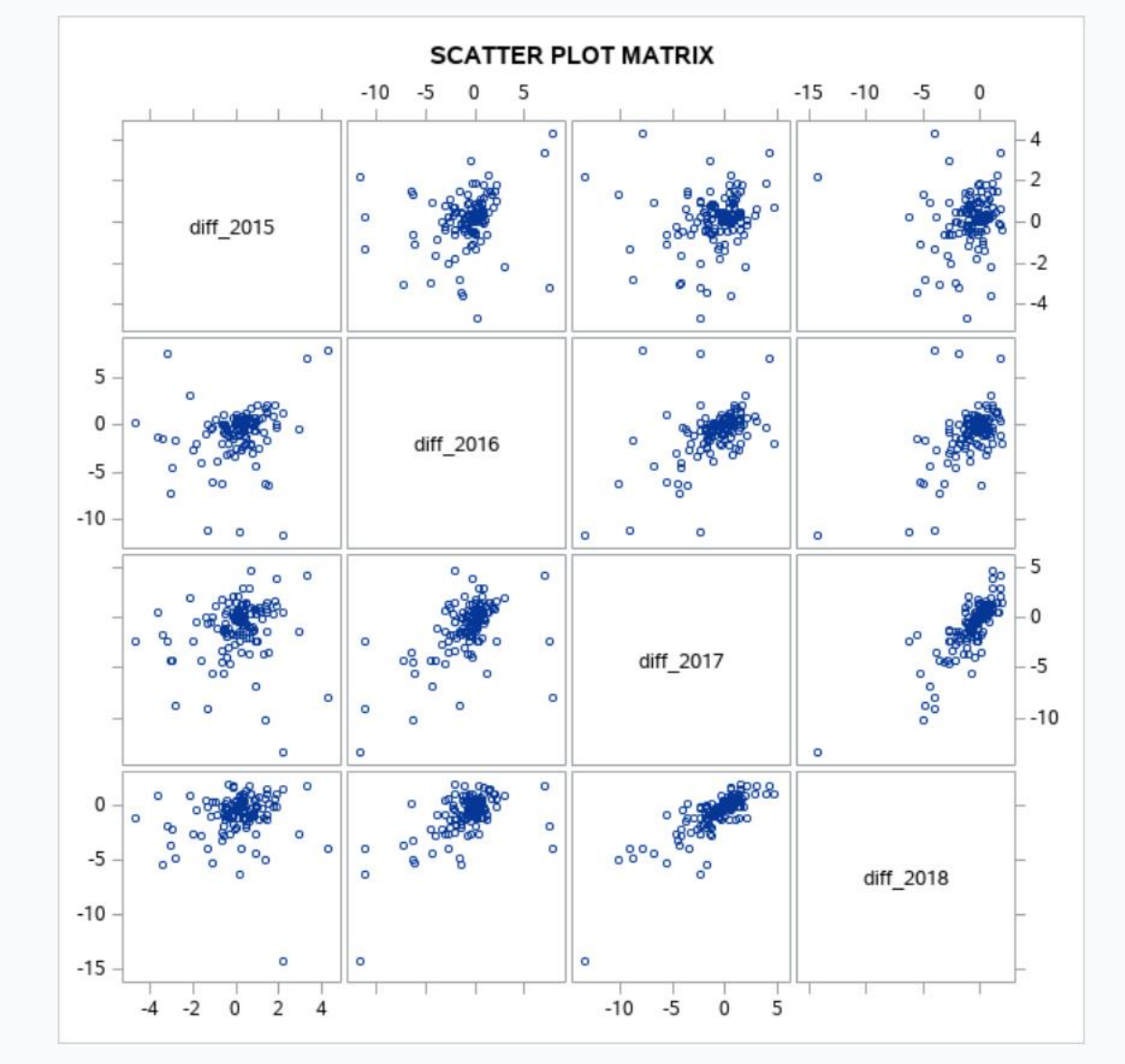

Figure 15 Scatter Plot of GEP Differences

One Line Conclusion:

There is a significant difference for the various GEP difference values per year but the overall projection is lower.

Part B)

```
PROC SQL;
CREATE TABLE WORK.QUERY3
AS
SELECT GEP1.Country_Code, GEP1.y_2001, GEP1.y_2002, GEP1.y_2003, GEP1.y_2004,
GEP1.y_2005, GEP1.y_2006, GEP1.y_2007, GEP1.y_2008, GEP1.y_2009, GEP1.y_2010,
GEP1.y_2011, GEP1.y_2012, GEP1.y_2013, GEP1.y_2014, GEP1.y_2015, GEP1.y_2016,
GEP1.y_2017, GEP1.y_2018, SUPPLEMENT._2019__2019_, SUPPLEMENT._2020__2020_
```

```
FROM WORK.GEP1 GEP1
LEFT JOIN WORK.SUPPLEMENT SUPPLEMENT
ON
   ( GEP1.Country_Name = SUPPLEMENT.Country_Name ) ;
QUIT;


PROC TRANSPOSE DATA = WORK.QUERY3 OUT = UPDATED;

DATA UPDATED1 (RENAME = (_NAME_ = YEAR COL1 = ARG COL2 = CHN COL3 = GRC COL4 =
GBR COL5 = USA ));
SET UPDATED;
RUN;

proc sgplot data= UPDATED1;
       series x=YEAR y=ARG/MARKERS;
       series x = YEAR y = USA/MARKERS;
       series x = YEAR y = CHN/MARKERS;
       series x = YEAR y = GRC/MARKERS;
       series x = YEAR y = GBR/MARKERS;
       xaxis grid display=(nolabel);
       yaxis grid label="GEP";
       TITLE "UPDATED GEP PLOT FOR FIVE SELECTED COUNTRIES";
run;
```

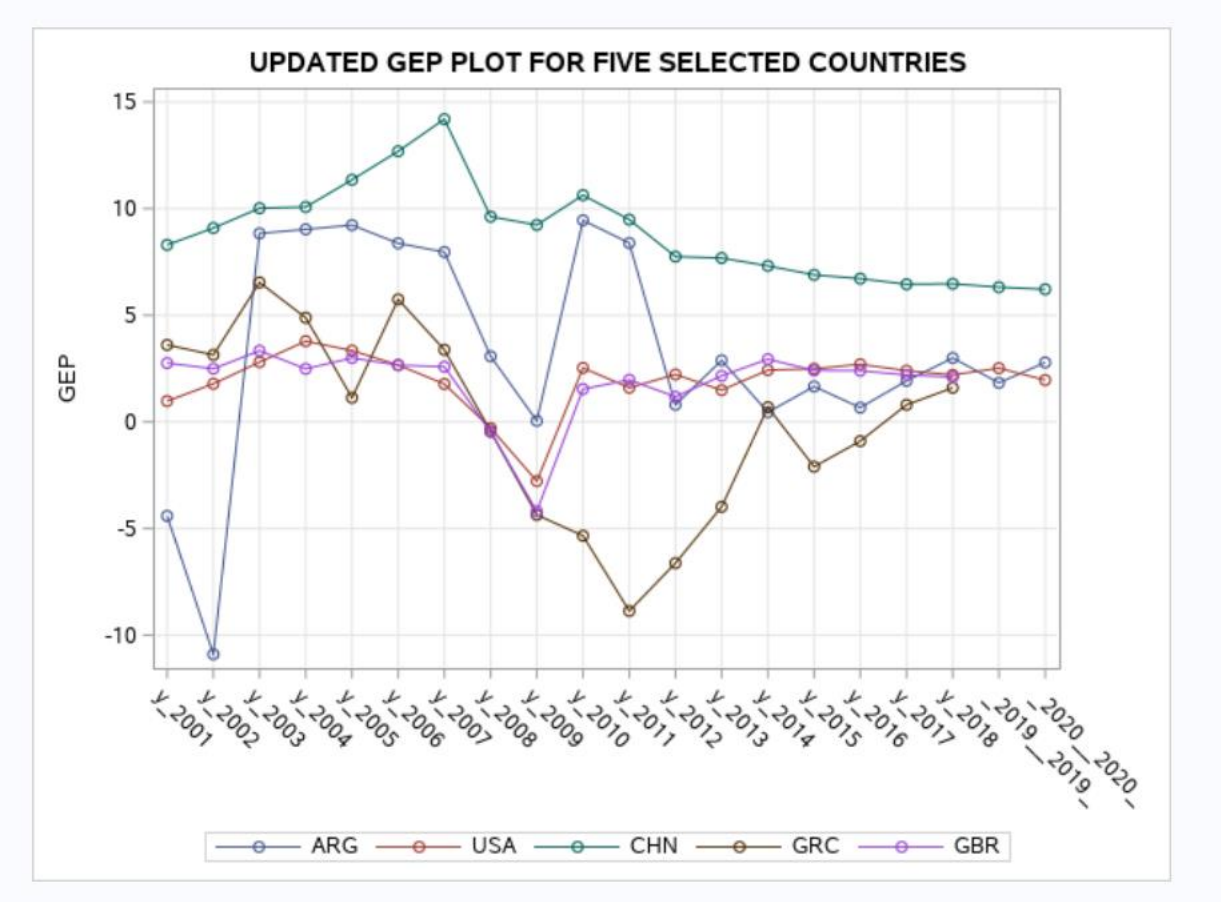

Figure 16 Updated GEP for Five Countries

Part C)

```
proc means data= QUERY3;
output out=GEP_Means;
run;
proc transpose data=GEP_Means out=GEP_Means_Trans;
id _STAT_;
run;


data GEP_Updated_Years;
set GEP_Updated_Means_Trans;
Variance=STD**2;
if _NAME_="_TYPE_" then delete;
if _NAME_="_FREQ_" then delete;
if _NAME_="y_2001" then Year=2001;
if _NAME_="y_2002" then Year=2002;
if _NAME_="y_2003" then Year=2003;
if _NAME_="y_2004" then Year=2004;
if _NAME_="y_2005" then Year=2005;
if _NAME_="y_2006" then Year=2006;
if _NAME_="y_2007" then Year=2007;
if _NAME_="y_2008" then Year=2008;
if _NAME_="y_2009" then Year=2009;
if _NAME_="y_2010" then Year=2010;
if _NAME_="y_2011" then Year=2011;
if _NAME_="y_2012" then Year=2012;
if _NAME_="y_2013" then Year=2013;
if _NAME_="y_2014" then Year=2014;
if _NAME_="y_2015" then Year=2015;
if _NAME_="y_2016" then Year=2016;
if _NAME_="y_2017" then Year=2017;
if _NAME_="y_2018" then Year=2018;
if _NAME_="y_2019" then Year=2019;
if _NAME_="y_2020" then Year=2020;
run;


proc sgplot data=GEP_Updated_Years;
reg y=Variance x=Year;
yaxis label="GEP Variance";
xaxis label="Year";
TITLE "VARIANCE BY YEAR";
run;
```

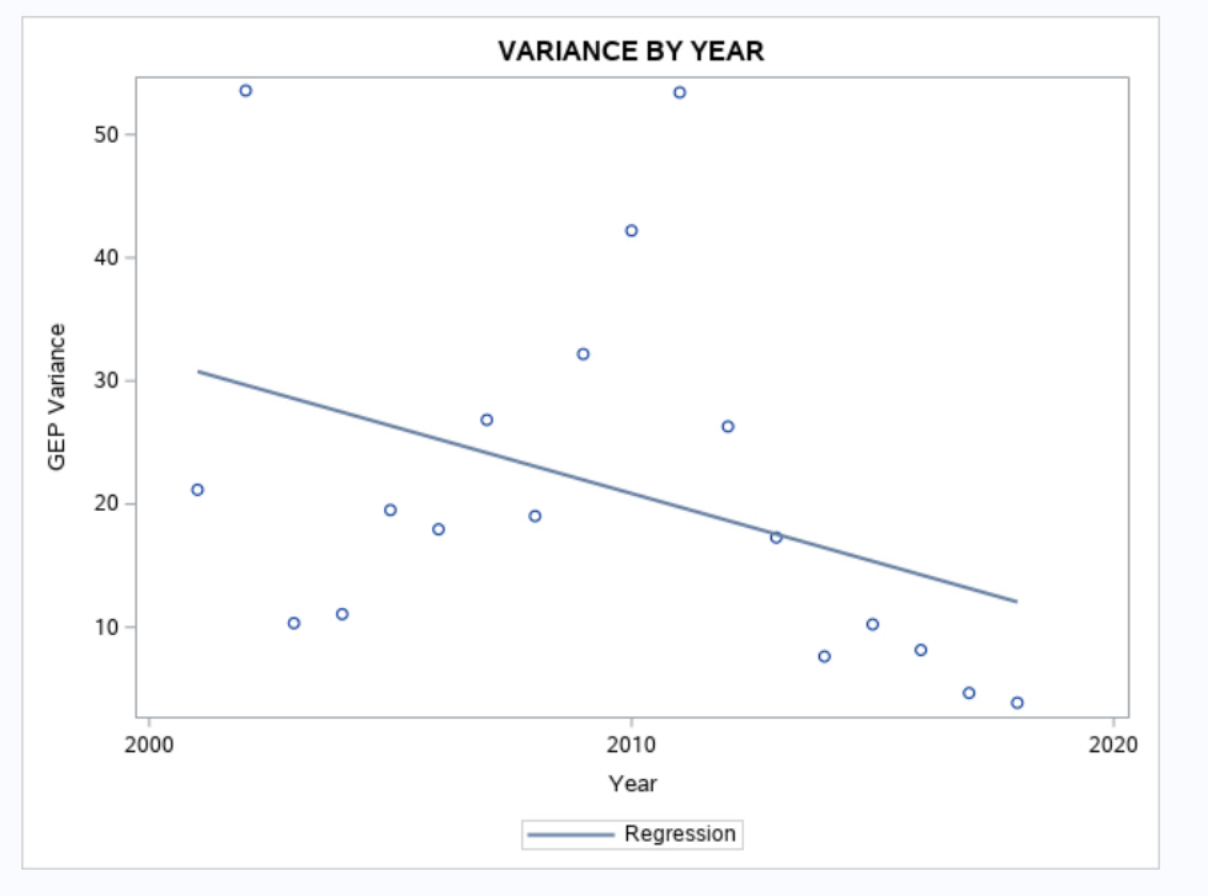

Figure 17 Variance By Year

Question 3

Part A)

```
PROC REG DATA = BOOKS PLOTS(MAXPOINTS = 10000);
MODEL average_rating = ratings_2 ;
```

The REG Procedure
Model: MODEL1
Dependent Variable: average_rating

Number of Observations Read	10000
Number of Observations Used	10000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8.69088	8.69088	136.07	<.0001
Error	9998	638.57782	0.06387		
Corrected Total	9999	647.26870			

Root MSE	0.25273	R-Square	0.0134
Dependent Mean	4.00219	Adj R-Sq	0.0133
Coeff Var	6.31470		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.01163	0.00265	1511.75	<.0001
ratings_2	1	-0.00000303	2.600964E-7	-11.66	<.0001

Figure 18 Simple Linear Regression Model Statistics

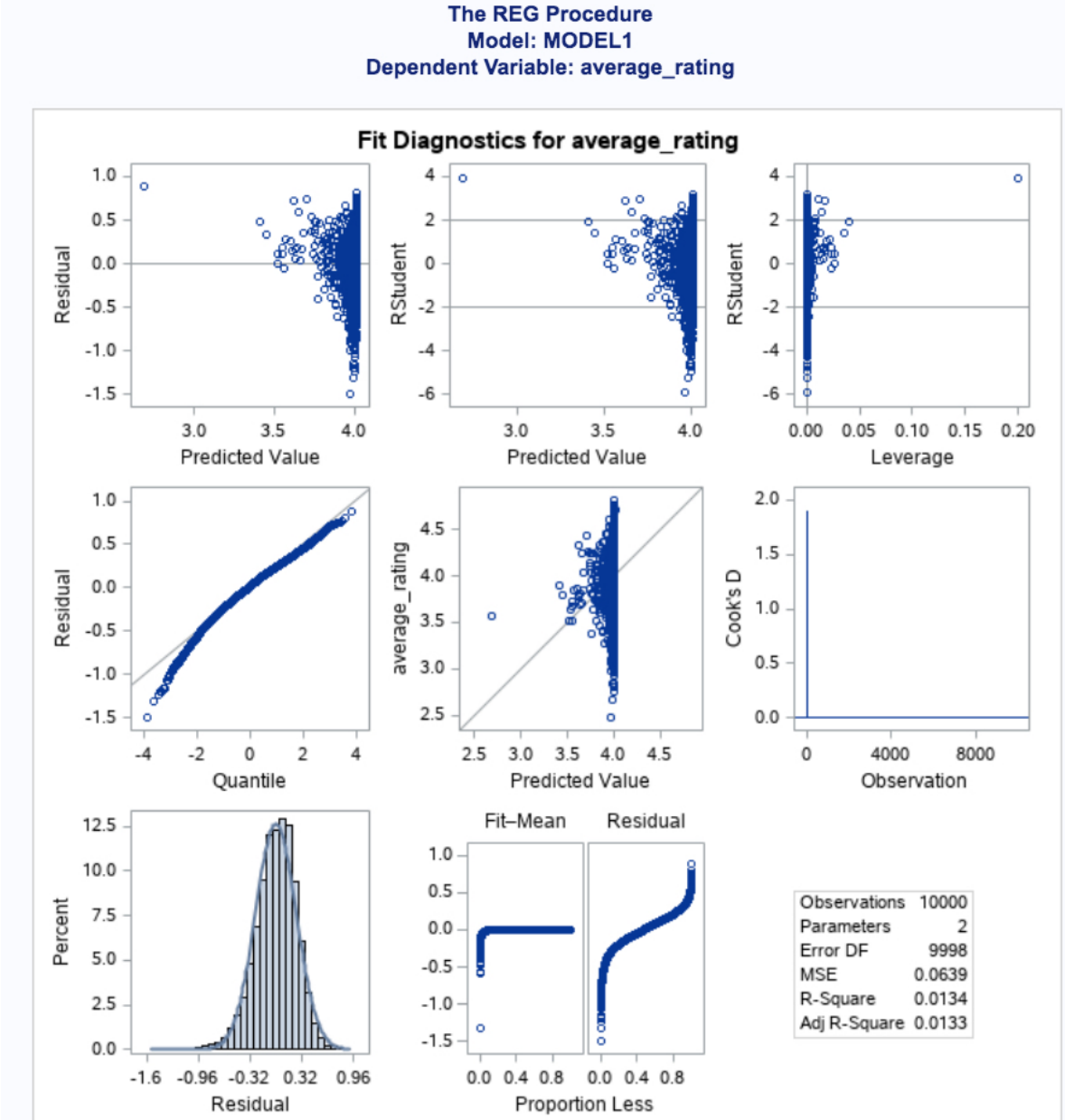

Figure 19 Summary of Plots for Simple Linear Regression

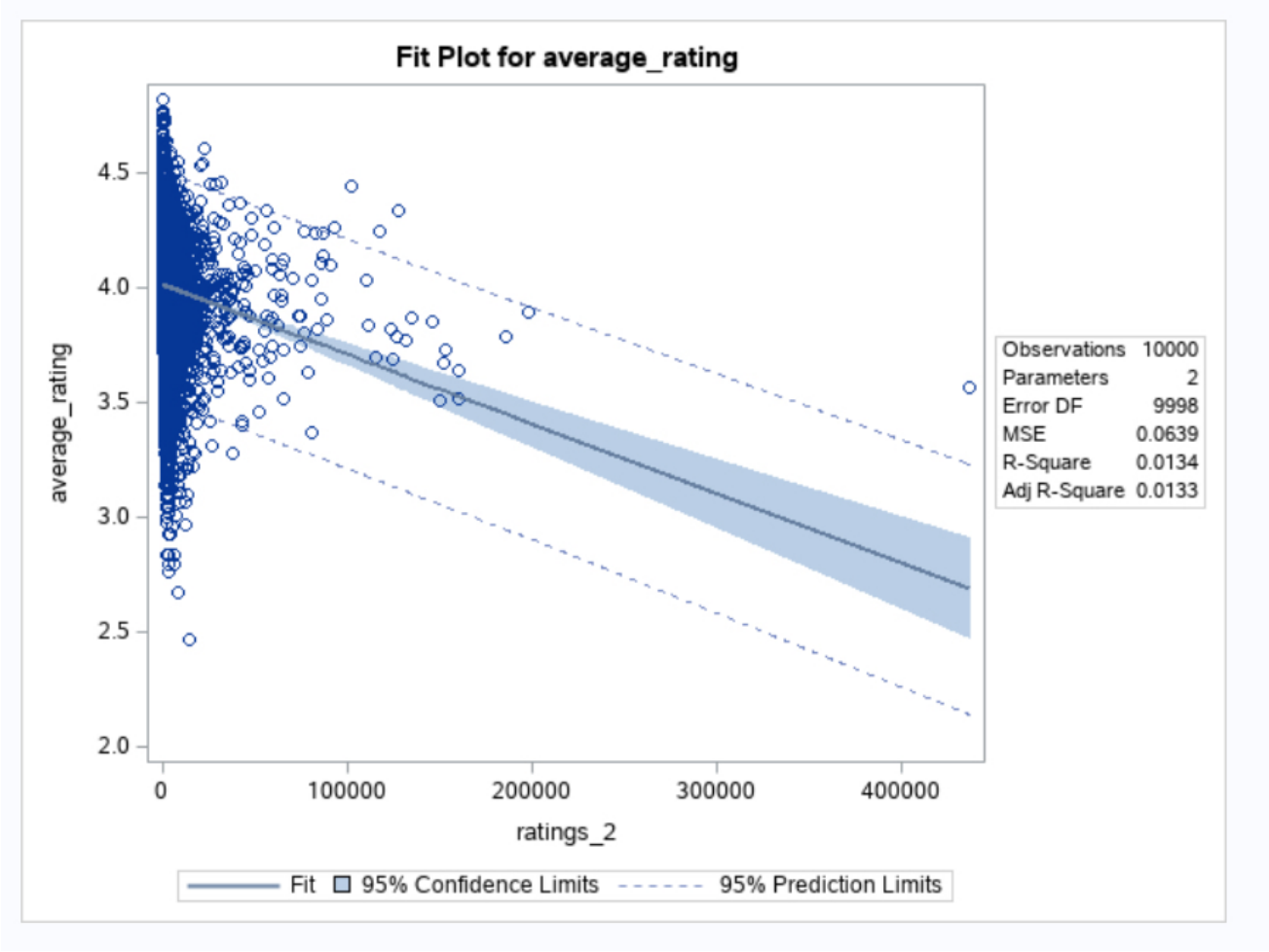

Figure 20 Simple Linear Regression Fit Plot

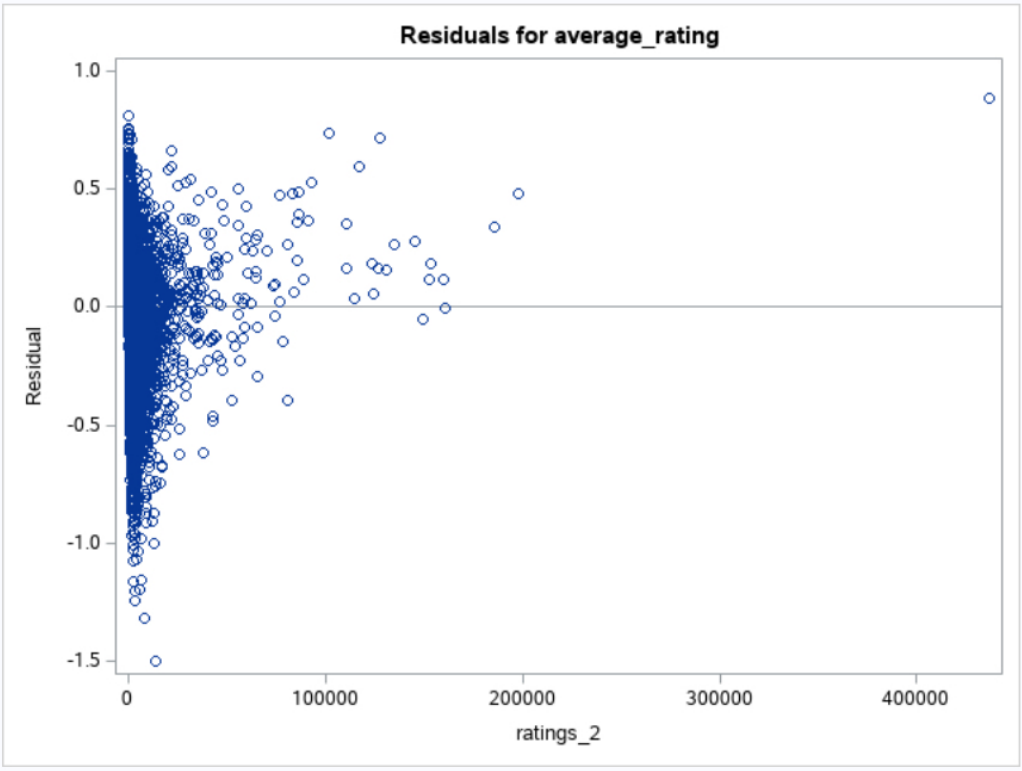

Figure 21 Simple Linear Regression Residuals Plot

Linear Regression Report

As shown in the linear regression output the R-Squared value is 0.0134 which represents a bad fitted model. Also shown is the residuals which are spread out and represent a poor model. So there is a little positive correlation between the average_rating and the ratings_2 variable.

```
PROC REG DATA = BOOKS PLOTS(MAXPOINTS = 10000);
MODEL average_rating = ratings_2 book_id;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: average_rating

Number of Observations Read	10000
Number of Observations Used	10000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	13.50773	6.75386	106.54	<.0001
Error	9997	633.76097	0.06340		
Corrected Total	9999	647.26870			

Root MSE	0.25178	R-Square	0.0209
Dependent Mean	4.00219	Adj R-Sq	0.0207
Coeff Var	6.29115		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.05474	0.00561	723.09	<.0001
ratings_2	1	-0.00000387	2.761596E-7	-14.00	<.0001
book_id	1	-0.00000810	9.295373E-7	-8.72	<.0001

Figure 22 Summary of Statistics for Multiple Linear Regression

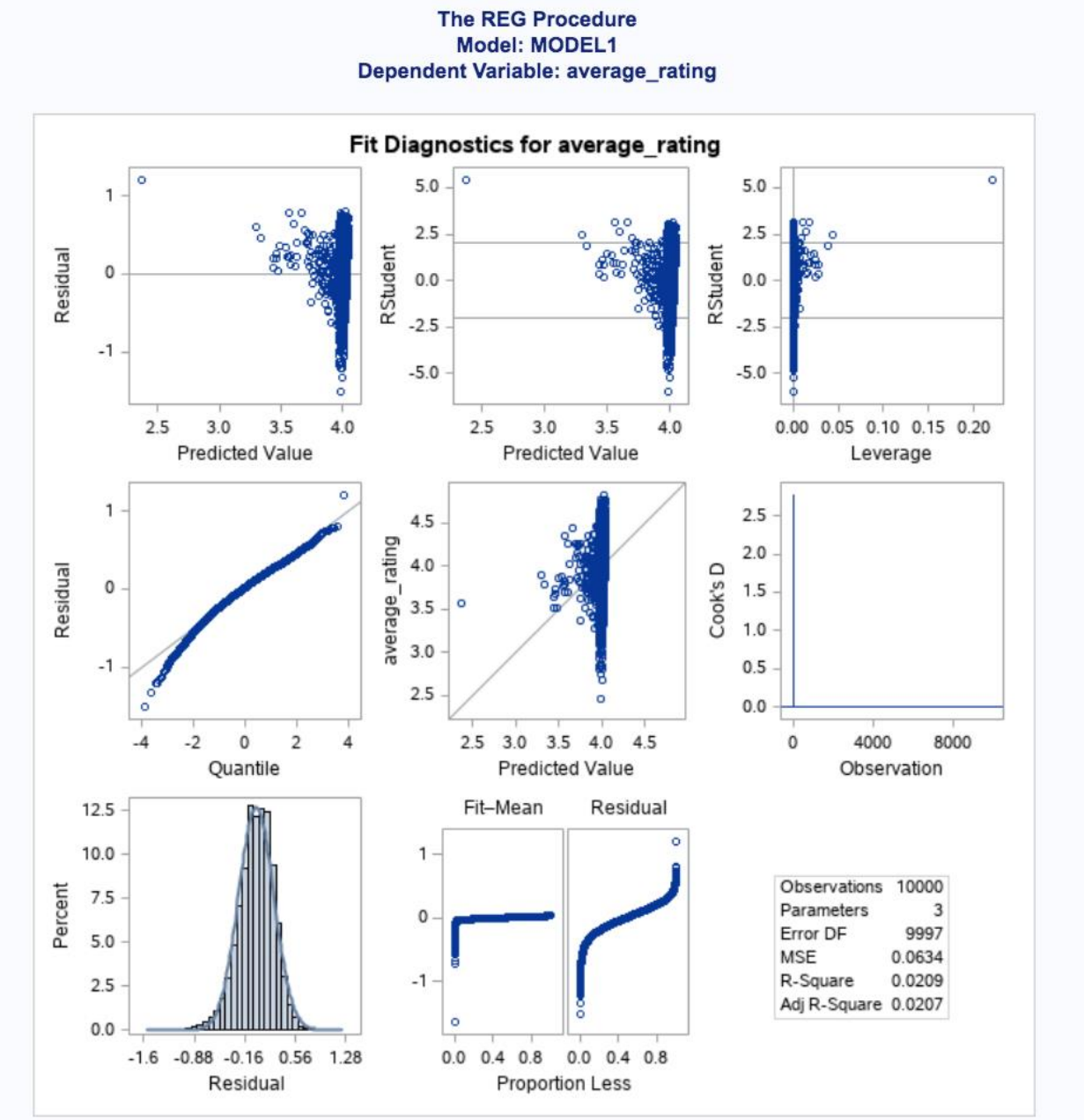

Figure 23 Summary Plots for Multiple Linear Regression

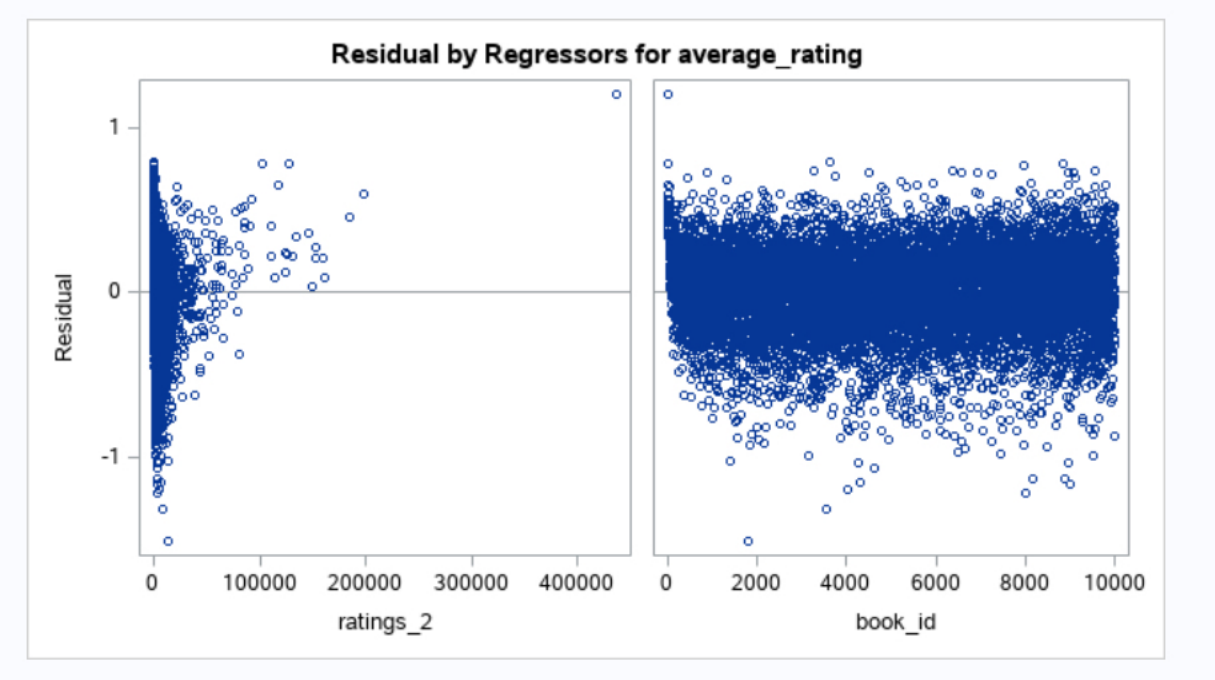

Figure 24 Residual Plot for Multiple Linear Regression

Multiple Linear Regression Report

For the multiple linear regression the R-Squared value is 0.0209 which represents a poorly fitted model. Also for the residuals they are spread out between -1 and 1 which also represents a poorly fitted model because the residuals should be close to zero.

Part B)

```
proc glm data=WORK.BOOKS PLOTS(MAXPOINTS = 10000);
      class language_code;
      model average_rating=language_code;
      means language_code / hovtest=levene welch plots=none;
      lsmeans language_code / adjust=tukey pdiff alpha=.05;
      run;
quit;
```

Class Level Information		
Class	Levels	Values
language_code	25	ara dan en en-CA en-GB en-US eng fil fre ger ind ita jpn mul nl nor per pol por rum rus spa swe tur vie

Number of Observations Read	10000
Number of Observations Used	8916

Dependent Variable: average_rating

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	24	4.9226647	0.2051110	3.20	<.0001
Error	8891	570.2002676	0.0641323		
Corrected Total	8915	575.1229323			

R-Square	Coeff Var	Root MSE	average_rating Mean
0.008559	6.327748	0.253244	4.002112

Source	DF	Type I SS	Mean Square	F Value	Pr > F
language_code	24	4.92266467	0.20511103	3.20	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
language_code	24	4.92266467	0.20511103	3.20	<.0001

Figure 25 ANOVA Table Statistics

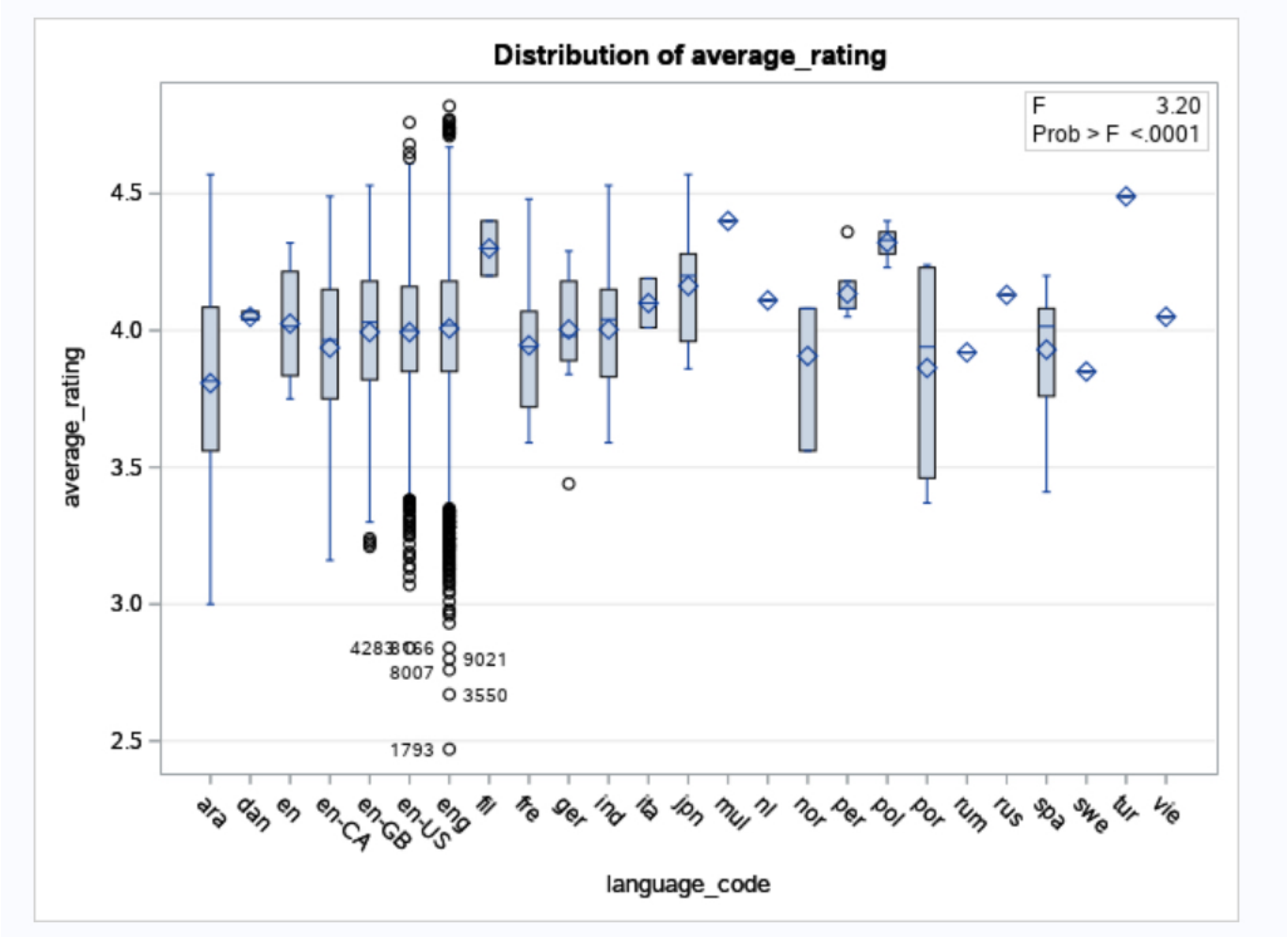

Figure 26 Distribution of Average Rating

Levene's Test for Homogeneity of average_rating Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
language_code	15	0.4815	0.0321	2.77	0.0003
Error	8889	102.9	0.0116		

Welch's ANOVA for average_rating			
Source	**DF**	**F Value**	**Pr > F**
language_code	17.0000	9.91	<.0001
Error	25.2898		

Figure 27 Levene's Test

Level of language_code	N	average_rating	
		Mean	Std Dev
ara	64	3.80750000	0.36349712
dan	3	4.05000000	0.01732051
en	4	4.02500000	0.24556058
en-CA	58	3.93724138	0.27393041
en-GB	257	3.99385214	0.26157341
en-US	2070	3.99358937	0.23956343
eng	6341	4.00746885	0.25601997
fil	2	4.30000000	0.14142136
fre	25	3.94560000	0.25426495
ger	13	4.00384615	0.22358960
ind	21	4.00380952	0.23986405
ita	2	4.10000000	0.12727922
jpn	7	4.16285714	0.23570563
mul	1	4.40000000	.
nl	1	4.11000000	.
nor	3	3.90666667	0.30022214
per	7	4.13428571	0.10768119
pol	6	4.32166667	0.06013873
por	6	3.86333333	0.37393404
rum	1	3.92000000	.
rus	1	4.13000000	.
spa	20	3.93000000	0.22461663
swe	1	3.85000000	.
tur	1	4.49000000	.
vie	1	4.05000000	.

Figure 28 Mean and Standard Deviation for Each Language Code

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

language_code	average_rating LSMEAN	LSMEAN Number
ara	3.80750000	1
dan	4.05000000	2
en	4.02500000	3
en-CA	3.93724138	4
en-GB	3.99385214	5
en-US	3.99358937	6
eng	4.00746885	7
fil	4.30000000	8
fre	3.94560000	9
ger	4.00384615	10
ind	4.00380952	11
ita	4.10000000	12
jpn	4.16285714	13
mul	4.40000000	14
nl	4.11000000	15
nor	3.90666667	16
per	4.13428571	17
pol	4.32166667	18
por	3.86333333	19
rum	3.92000000	20
rus	4.13000000	21
spa	3.93000000	22
swe	3.85000000	23
tur	4.49000000	24
vie	4.05000000	25

Figure 29 Least Squares Means for Each Language Code

Least Squares Means for effect language_code
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: average_rating

i/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1		0.9968	0.9953	0.4360	<.0001	<.0001	<.0001	0.5299	0.8247	0.6586	0.2576	0.9971	0.0771	0.8190	1.0000	1.0000	0.1741	0.0006	1.0000	1.0000	0.9999	0.9762	1.0000	0.5576	1.0000
2	0.9968		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9988	1.0000	1.0000	1.0000	1.0000	1.0000	0.9990	1.0000
3	0.9953	1.0000		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000	1.0000	0.9853	1.0000	1.0000	1.0000	1.0000	1.0000	0.9961	1.0000
4	0.4360	1.0000	1.0000		0.9985	0.9951	0.9241	0.9565	1.0000	1.0000	1.0000	1.0000	0.8714	0.9856	1.0000	1.0000	0.9666	0.0736	1.0000	1.0000	1.0000	1.0000	1.0000	0.9001	1.0000
5	<.0001	1.0000	1.0000	0.9985		1.0000	1.0000	0.9936	1.0000	1.0000	1.0000	1.0000	0.9913	0.9973	1.0000	1.0000	0.9994	0.2279	1.0000	1.0000	1.0000	1.0000	1.0000	0.9644	1.0000
6	<.0001	1.0000	1.0000	0.9951	1.0000		0.8997	0.9932	1.0000	1.0000	1.0000	1.0000	0.9897	0.9972	1.0000	1.0000	0.9993	0.2095	0.9999	1.0000	1.0000	1.0000	1.0000	0.9635	1.0000
7	<.0001	1.0000	1.0000	0.9241	1.0000	0.8997		0.9964	1.0000	1.0000	1.0000	1.0000	0.9968	0.9984	1.0000	1.0000	0.9999	0.2852	0.9997	1.0000	1.0000	0.9998	1.0000	0.9736	1.0000
8	0.5299	1.0000	1.0000	0.9565	0.9936	0.9932	0.9964		0.9737	0.9985	0.9978	1.0000	1.0000	1.0000	1.0000	0.9937	1.0000	1.0000	0.9207	1.0000	1.0000	0.9614	0.9994	1.0000	1.0000
9	0.8247	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9737		1.0000	1.0000	1.0000	0.9529	0.9901	1.0000	1.0000	0.9913	0.1629	1.0000	1.0000	1.0000	1.0000	1.0000	0.9221	1.0000
10	0.6586	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	1.0000		1.0000	1.0000	0.9998	0.9989	1.0000	1.0000	1.0000	0.6632	1.0000	1.0000	1.0000	1.0000	1.0000	0.9814	1.0000
11	0.2576	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9978	1.0000	1.0000		1.0000	0.9995	0.9987	1.0000	1.0000	1.0000	0.5274	1.0000	1.0000	1.0000	1.0000	1.0000	0.9780	1.0000
12	0.9971	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	1.0000
13	0.0771	1.0000	1.0000	0.8714	0.9913	0.9897	0.9968	1.0000	0.9529	0.9998	0.9995	1.0000		1.0000	1.0000	0.9993	1.0000	1.0000	0.9155	1.0000	1.0000	0.9271	1.0000	1.0000	1.0000
14	0.8190	1.0000	0.9999	0.9856	0.9973	0.9972	0.9984	1.0000	0.9901	0.9989	0.9987	1.0000	1.0000		1.0000	0.9944	1.0000	1.0000	0.9631	0.9998	1.0000	0.9857	0.9986	1.0000	1.0000
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9937	1.0000	1.0000	1.0000	1.0000	0.9993	0.9944	1.0000		0.9999	0.8215	1.0000	1.0000	1.0000	1.0000	1.0000	0.9557	1.0000
17	0.1741	1.0000	1.0000	0.9666	0.9994	0.9993	0.9999	1.0000	0.9913	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999		0.9999	0.9705	1.0000	1.0000	0.9829	1.0000	0.9999	1.0000
18	0.0006	0.9988	0.9853	0.0736	0.2279	0.2095	0.2852	1.0000	0.1629	0.6632	0.5274	1.0000	1.0000	1.0000	1.0000	0.8215	0.9999		0.2277	0.9993	1.0000	0.1399	0.9924	1.0000	1.0000
19	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9207	1.0000	1.0000	1.0000	1.0000	0.9155	0.9631	1.0000	1.0000	0.9705	0.2277		1.0000	1.0000	1.0000	1.0000	0.8370	1.0000
20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	1.0000	0.9993	1.0000		1.0000	1.0000	1.0000	0.9976	1.0000
21	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000	1.0000
22	0.9762	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9614	1.0000	1.0000	1.0000	1.0000	0.9271	0.9857	1.0000	1.0000	0.9829	0.1399	1.0000	1.0000	1.0000		1.0000	0.9027	1.0000
23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9994	1.0000	1.0000	1.0000	1.0000	1.0000	0.9986	1.0000	1.0000	1.0000	0.9924	1.0000	1.0000	1.0000	1.0000		0.9879	1.0000
24	0.5576	0.9990	0.9961	0.9001	0.9644	0.9635	0.9736	1.0000	0.9221	0.9814	0.9780	0.9999	1.0000	1.0000	1.0000	0.9557	0.9999	1.0000	0.8370	0.9976	1.0000	0.9027	0.9879		1.0000
25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	

Figure 30 Least Squares Matrix

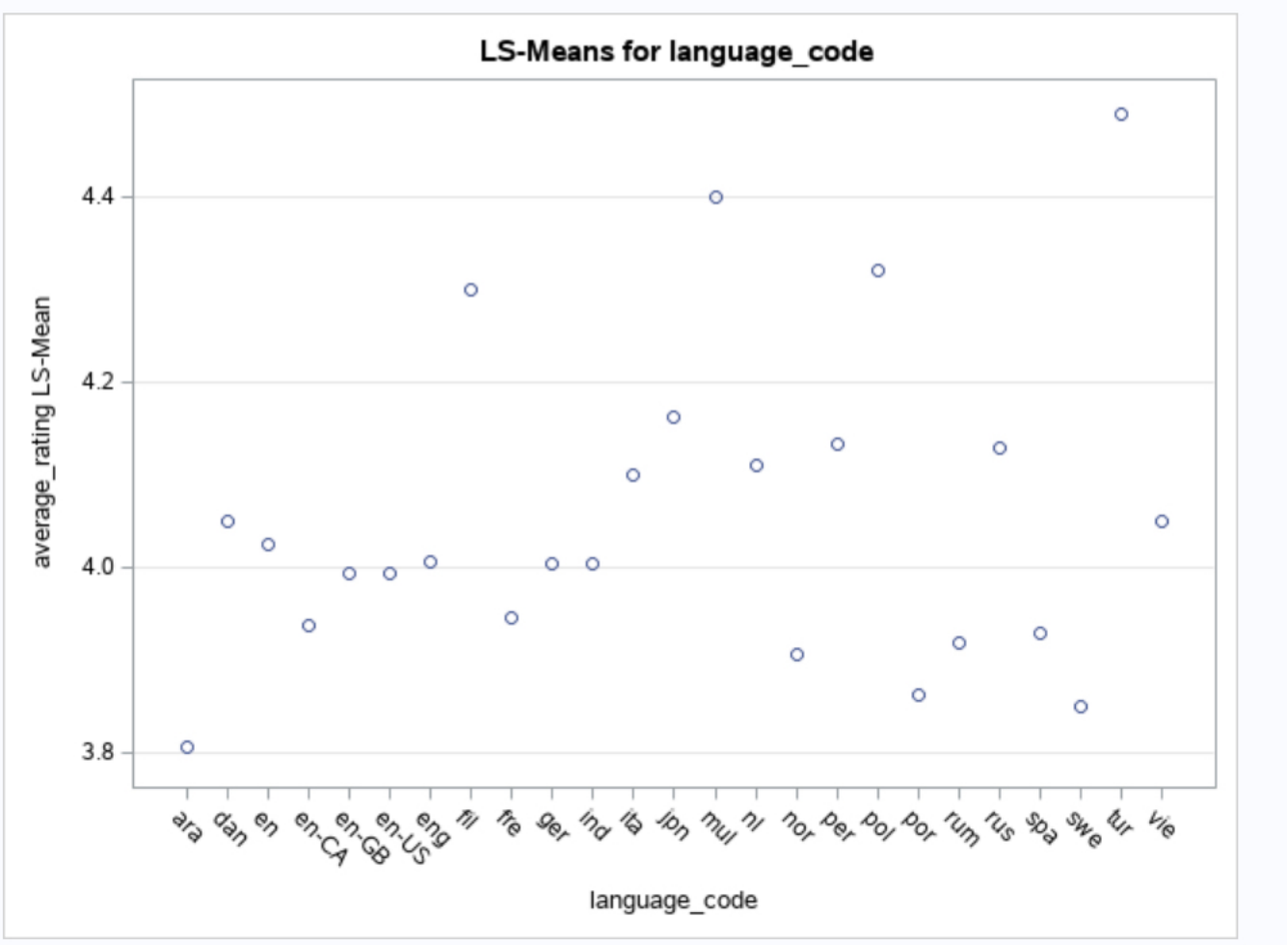

Figure 31 Least Squares Means Scatter Plot

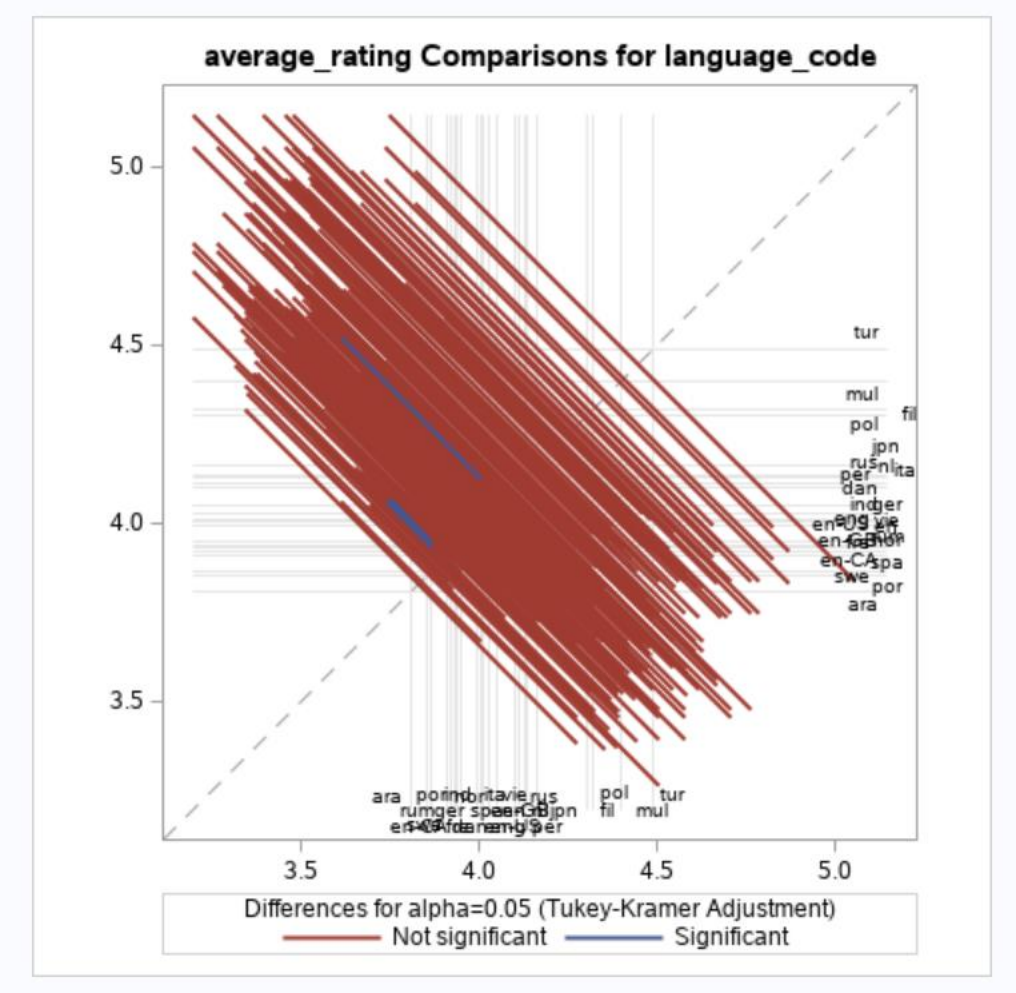

Figure 32 Plot of Language Codes

Looking at the figure 28 and 29 I see that the least squares means of the average ratings are all around a value of 4 with respect to the language code.

From figure 26 the distribution is well balanced being around the average value of 4 with very few outliers occurring at around a value of 4.4.

In conclusion, the average ratings are all near an equal mean of 4 given the language_code.