Neeraj Ankam
Ne981078

Assignment 3

Due Dec 1st

Provide the SAS statements for each answer and where a PROC statement is used, a screenshot of the
result, or the pdf output of the result from SAS's options.
Any difficulties, are to be seen by the datamining lab in the maths building, 150D/E, where a GTA
member can assist you from 9am till 4pm on weekdays.
This work is to be done individually and submitted through webcourses. Some aspects of the questions
require a bit of investigation on how to implement the requests within SAS.
There are two datasets GDP.csv and GEP.csv . The files contain country names and numerical data.
They come from the worldbank databank. The GDP is the gross domestic product for each country and
the GEP is the Global Economic Prospect for each country over a set of years in the past's prediction
and into the future. The headers have been removed but for the GDP.csv they are:
Country code, Country name, gdp
for GEP.csv:
Country Name, CountryCode, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011,
2012, 2013, 2014, 2015, 2016, 2017, 2018

Question 1 40pnts
a) Load in these two files and print them. (hint: due to the quotation marks around multiple worded
country names, the flag 'dsd' with the infile statement can be useful) (5pnts)

```
data assignment;
infile 'C:\Users\neeraj ankam\Downloads\GDP.csv' dlm=',' missover dsd;
format Country_Code $ 3. CountryName $ 50.  GDP comma8.0;
input Country_Code $ CountryName $ GDP;
run;

proc print data=assignment;
run;

data at;
infile 'C:\GEP_Data.csv' dlm=',' missover dsd;
format CountryName $ 50. CountryCode $ 3.;
input CountryName $ CountryCode $ y2001 y2002 y2003 y2004 y2005 y2006 y2007
y2008 y2009 y2010 y2011 y2012 y2013 y2014 y2015 y2016 y2017 y2018;
run;

proc print data=at;
run;
```

The SAS System

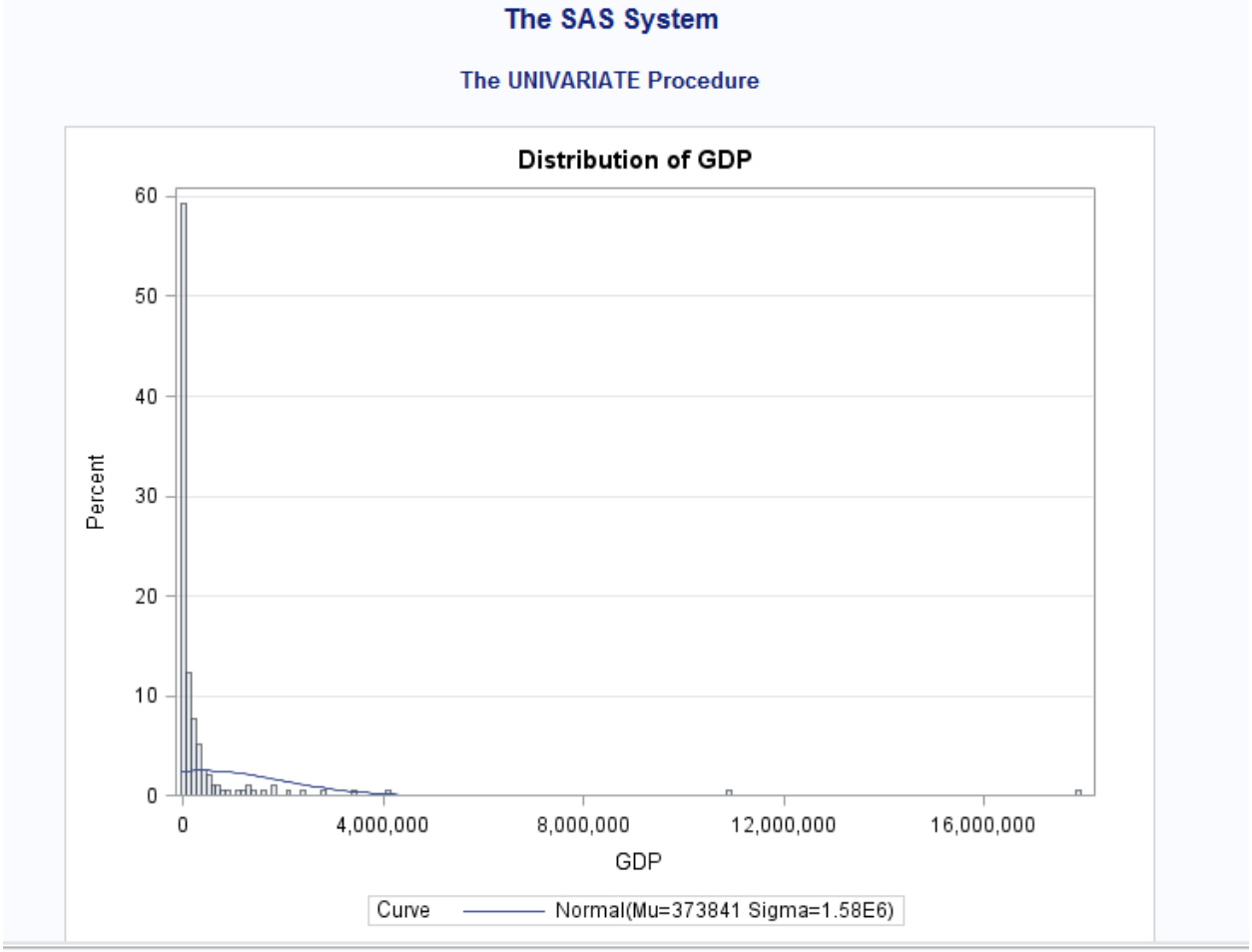
Obs	Country_Code	CountryName	GDP
1	USA	United States	17946996
2	CHN	China	10866444
3	JPN	Japan	4123258
4	DEU	Germany	3355772
5	GBR	United Kingdom	2848755
6	FRA	France	2421682
7	IND	India	2073543
8	ITA	Italy	1814763
9	BRA	Brazil	1774725
10	CAN	Canada	1550537
11	KOR	Korea, Rep.	1377873

The SAS System

Obs	CountryName	CountryCode	y2001	y2002	y2003	y2004	y2005	y2006	y2007	y2008	y2009	y2010	y2011	y2012	y2013	y2014
1	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
2	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
3	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
4	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
5	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
6	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
7	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
8	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
9	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
10	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
11	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
12	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
13	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
14	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
15	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

b) Produce a histogram for the GDP values in the dataset. (hint: proc univariate data=data1; histogram;
run;) This can get you started and if you look at the documentation, there are ways to configure the
histogram to change its appearance and bins
(http://support.sas.com/documentation/cdl/en/procstat/66703/HTML/default/viewer.htm#procstat_univ
ariate_syntax09.htm) any extra configuration earns more points. (3pnts).

```
proc univariate data=assignment;
var GDP;
histogram/ normal midpoints = 1000000 to 17000000 by 100000 ctext= blue grid
wbarline=5 waxis=5;
run;
```



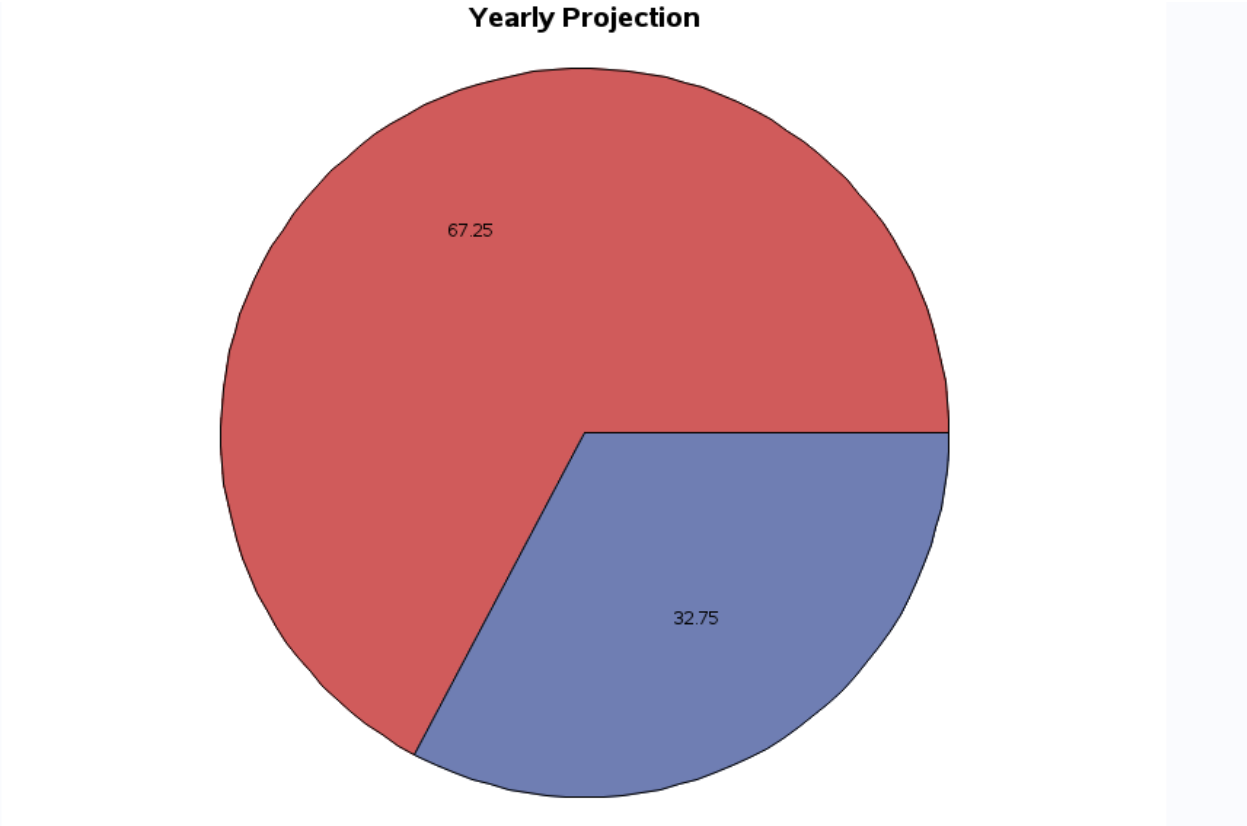
c) Produce a Pie Chart (or chart of your choice) of the total GDP values, the percentage of GDP produced by the top 10 countries. (4pnts)

```
data pieChart;
infile 'S:\piechart.csv'
delimiter=','
DSD;

input  status $  gdp ;
run;

proc print data  = pieChart;
run;

proc gchart data=pieChart;
pie status / sumvar=gdp
other=0 descending legend=legend1 value=inside coutline=black noheading;
run;
quit;
```



d) Use PROC SQL to select from GEP the countries (United States,USA), (Greece,GRC), (China,CHN), (United Kingdom,UK), (Argentina,ARG) and then print. (4pnts)

```
libname sql 'Work';
proc sql;
create table GEP as
select *
from Work.at
where CountryCode='USA' or CountryCode='GRC' or CountryCode='CHN' or
CountryCode='GBR' or CountryCode='ARG';

select * from GEP;
```

The SAS System

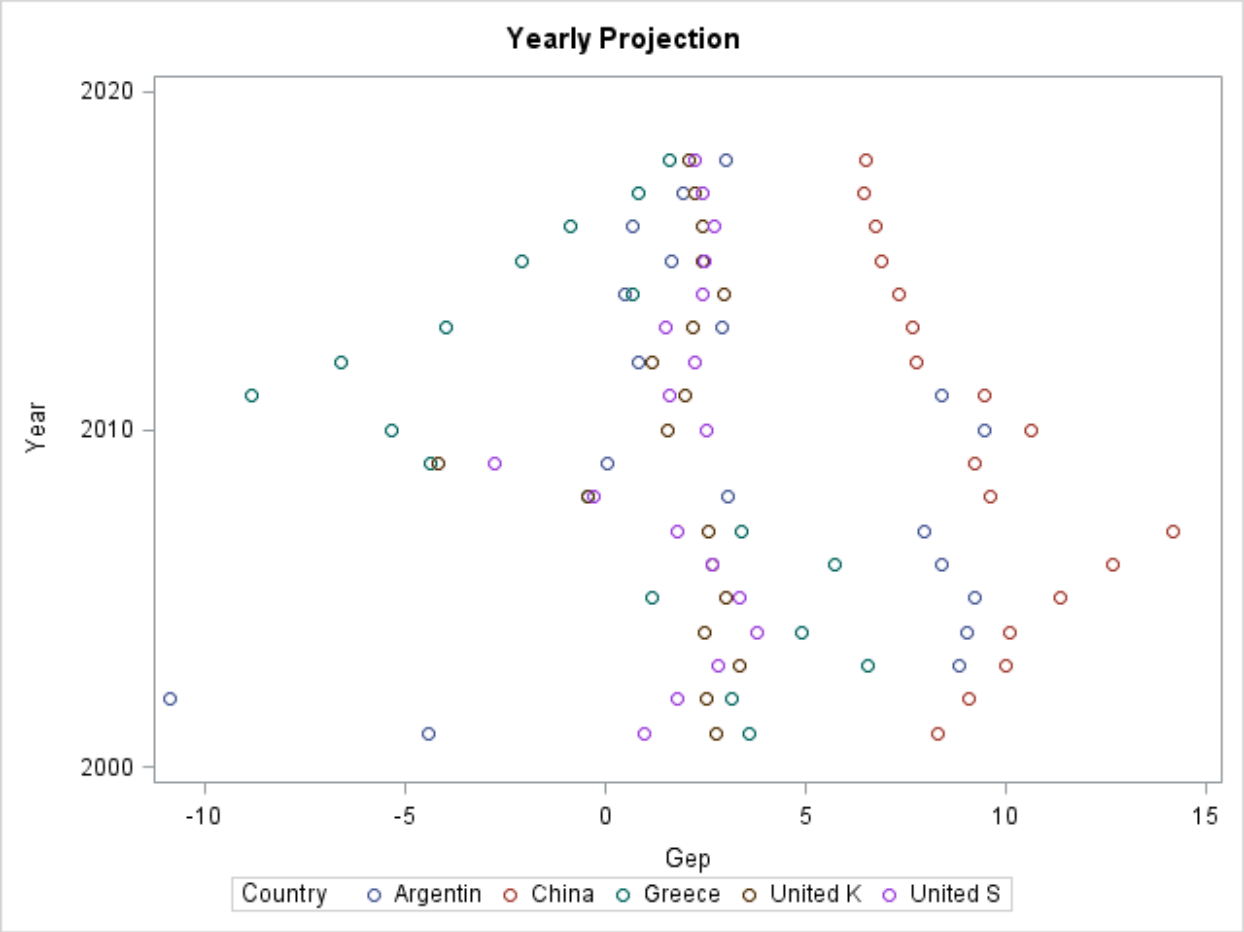
CountryName	CountryCode	y2001	y2002	y2003	y2004	y2005	y2006	y2007	y2008	y2009	y2010	y2011	y2012	y2013	y201
Argentina	ARG	-4.40884	-10.8945	8.837041	9.029573	9.226321	8.375248	7.965573	3.074945	0.050024	9.451578	8.386451	0.801761	2.885351	0.45360
China	CHN	8.298374	9.090909	10.01997	10.07564	11.35239	12.68823	14.19496	9.623377	9.233546	10.63171	9.484511	7.750294	7.68382	7.31643
Greece	GRC	3.609537	3.14236	6.537896	4.884141	1.133032	5.74489	3.382163	-0.43795	-4.36041	-5.33692	-8.86572	-6.62053	-3.98094	0.69027
United Kingdom	GBR	2.757963	2.493987	3.336665	2.488391	2.996425	2.661759	2.5861	-0.46688	-4.19194	1.540177	1.972399	1.179056	2.159904	2.94019
United States	USA	0.976146	1.787545	2.805816	3.78567	3.346281	2.665374	1.77914	-0.29112	-2.77604	2.532111	1.601066	2.224279	1.489446	2.42758

d) From the data in the previous step, plot the data for these country yearly projections. (5pnts)

```
data yearlyProjection;
infile 'S:\YearlyProjection.csv' delimiter=',' missover DSD;
input Country $ Year Gep;
run;
```

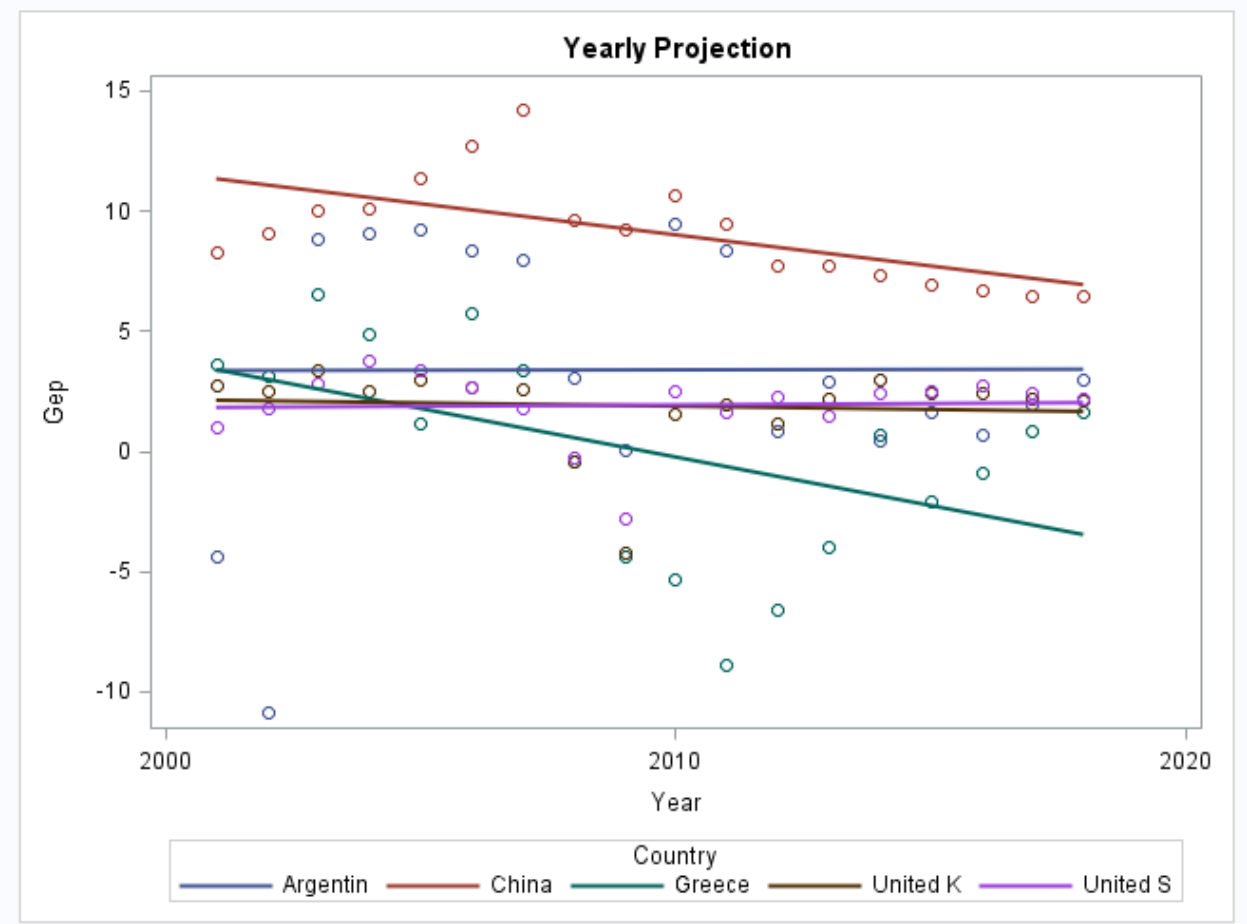
Neeraj Ankam
Ne981078

```
PROC sgscatter  DATA=yearlyProjection;
   PLOT year * gep
   / group = country;
   title 'Yearly Projection';
RUN;
```



e) Use 'PROC SGPLOT data= ; reg x= y= ; run;' or any other suitable method you wish to plot a line
of through the data points over these years for these countries. (any method will suffice, even if the
lines are on different plots; if you choose to use a subset of the years a single point will be removed)(5pnts)

```
PROC SGPLOT data=yearlyProjection;
reg x=Year y=Gep/group = Country;
run;
```



f) Use PROC SQL to join these two tables together. (3pnts)

```
proc sql;
create table joinedtable
as select *
from at join assignment on at.CountryCode = assignment.Country_Code;
select * from joinedtable;
```

```
661   proc sql;
662   create table joinedtable
663   as select *
664   from at join assignment on at.CountryCode = assignment.Country_Cod
WARNING: Variable CountryName already exists on file WORK.JOINEDTABLE.
NOTE: Table WORK.JOINEDTABLE created, with 178 rows and 22 columns.
```

Yearly Projection

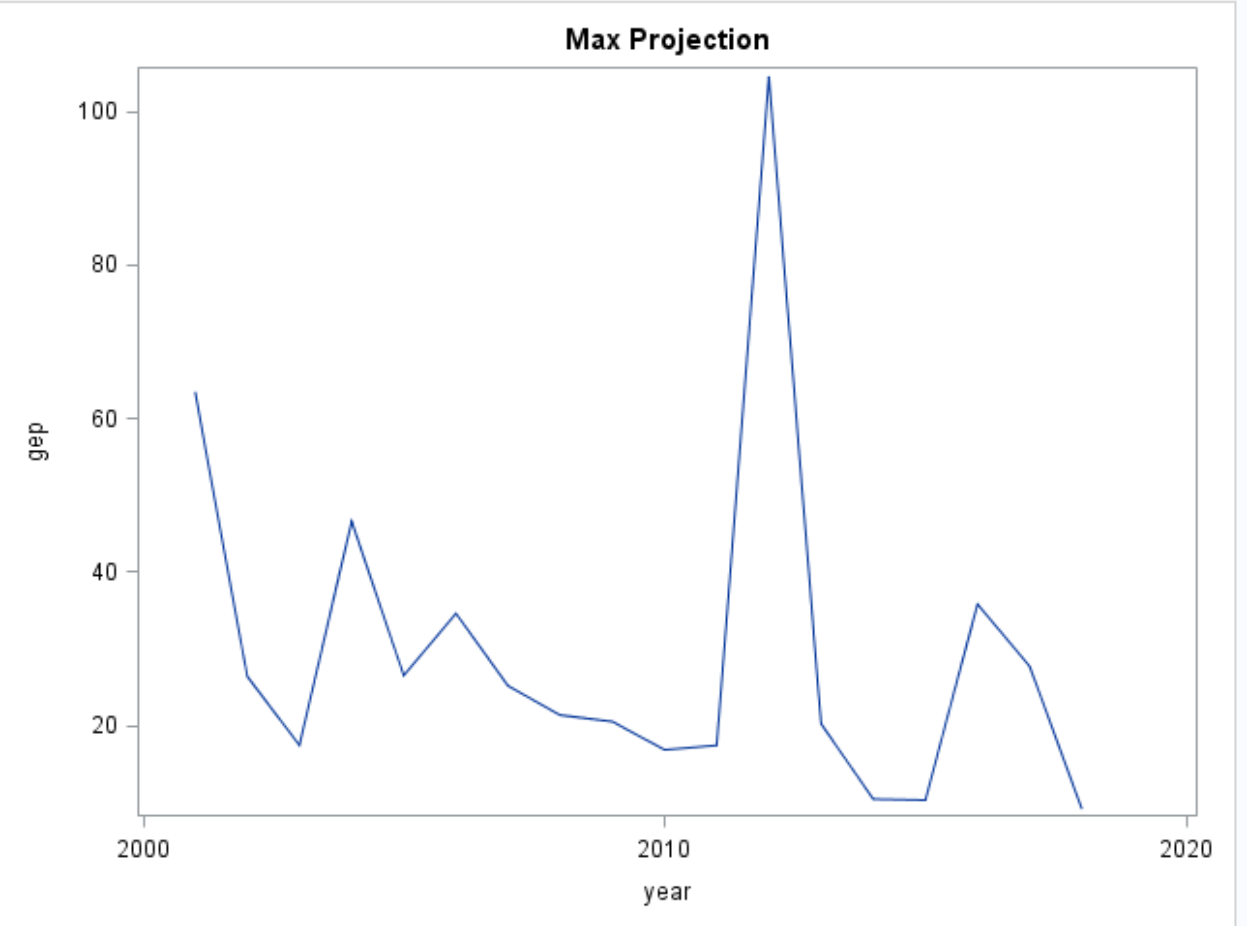
CountryName	CountryCode	y2001	y2002	y2003	y2004	y2005	y2006	y2007	y2008	y2009	y2010	y2011	y2012	y2013	y201
Afghanistan	AFG	20.06408	16.71114	14.31838	9.438758	14.51572	11.18704	11.13201	3.4	20.4	8.4	6.1	14.4	2	1.
Albania	ALB	7	2.9	5.7	5.9	5.499912	5.00001	5.900104	7.536299	3.354239	3.706906	2.545359	1.623699	1.417423	2.0
Algeria	DZA	4.612524	5.6	7.2	4.3	5.984991	1.7	3.4	2.4	1.59998	3.60017	2.9	3.4	2.8	3.
Angola	AGO	4.220965	13.82179	5.24757	10.87947	18.26147	20.73512	22.59305	13.81715	2.41287	3.407655	3.918597	5.155441	6.800058	3.90126
Antigua and Barbuda	ATG	-3.19424	2.924251	5.913631	5.28642	6.083228	13.3764	9.498944	0.071111	-12.036	-7.14299	-1.79344	4.019542	-0.07136	3.22150
Argentina	ARG	-4.40884	-10.8945	8.837041	9.029573	9.226321	8.375248	7.965573	3.074945	0.050024	9.451578	8.386451	0.801761	2.885351	0.45360
Armenia	ARM	9.556641	13.1863	14.0408	10.46784	13.86571	13.198	13.7492	6.9	-14.15	2.2	4.7	7.2	3.3	3.

y2013	y2014	y2015	y2016	y2017	y2018	Country_Code	GDP
2	1.3	1.9	3.1	3.9	5	AFG	19,199
1.417423	2.02	2.7	3.4	3.5	3.5	ALB	11,456
2.8	3.8	2.8	3.9	4	3.8	DZA	166,839
6.800058	3.901265	3.037772	3.306046	3.816762	3.831209	AGO	102,643
-0.07136	3.221505	2.0111	2.42237	2.73491	2.73491	ATG	1,297
2.885351	0.453604	1.659752	0.669259	1.926797	2.996408	ARG	583,169
3.3	3.5	2.5	2.2	2.8	3	ARM	10,561

g) Using the dataset from f) how is the maximum projection from each year trending? (eg. for each year what is the max projection, and draw a line to show it) (4pnts)

```
data projectionChart;
infile 'S:\maximum_projection.csv'  delimiter=',' DSD;
input  gep  year ;
run;
proc sgplot data=projectionChart;
series x=year y=gep;
run;
```

The data ‘projectionChart’ has maximum gep for each year which has been plotted in the graph.



Yearly Projection

max2001	max2002	max2003	max2004	max2005	max2006	max2007	max2008	max2009	max2010	max2011	max2012	max2013	max2014	max2015
63.37988	26.26858	17.32	46.49997	26.4	34.5	25.049	21.24564	20.4	16.73156	17.29078	104.5	20.12359	10.3	10.16779

max2013	max2014	max2015	max2016	max2017	max2018
20.12359	10.3	10.16779	35.7	2017	2018

h) Use PROC CORR with the country GDP and any year of choice in the projection (GEP). (2pnts)

```
proc corr data=joinedtable;
var GDP y2017;
run;
```

Yearly Projection

The CORR Procedure

2 Variables:	GDP y2017

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
GDP	178	405686	1648145	72212118	38.00000	17946996
y2017	178	4.01059	2.68540	713.88447	-0.40463	27.60000

Pearson Correlation Coefficients, N = 178
Prob > |r| under H0: Rho=0

	GDP	y2017
GDP	1.00000	-0.07081 0.3476
y2017	-0.07081 0.3476	1.00000

I) What conclusions can you draw from these results? (5pnts)

In the above correlation matrix, I have checked for the correlation between two variables GDP and 2017 year. The N value shows the total number of non-missing attributes for correlation and all the values are present. Values like the Mean, Sum et cetera of the selected variables has been described above.

The correlation value is negative and hence there is a negative correlation between GDP and GEP 0f 2017 Year. Since the value is small, it isn't a strong negative correlation. Since the p-value is >>> 0.05, we can safely say that the results are statistically insignificant and there is weak evidence against the Null hypothesis.