Assignment - 4

SAS Code

```
LIBNAME mydata "/courses/d1406ae5ba27fe300" access=readonly;
DATA new; set mydata.gapminder;
LABEL incomeperperson = "Gross Domestic Product per capita in constant 2000 US dollar (2010)"
      employrate = "% of total population, age above 15, that has been employed during 2007"
      urbanrate = "Uban population (% of total population) (2008)"
      lifeexpectancy = "Life expectancy at birth (years) (2011)"

      incpp_cat="Categorical disribution of Gross Domestic Product per capita in constant 2000 US dollar (2010)"
      emprate_cat="Categorical disribution of % of total population, age above 15, that has been employed during 2007"
      urate_cat = "Categorical distribution of Uban population (% of total) (2008)"
      lexp_cat = "Categorical disribution of Life expectancy at birth (years) (2011)";


/*making categories based on income per person for a country*/
/*IF incomeperperson = . THEN incpp_cat="Data NA       ";
else*/ IF incomeperperson LT 1000 THEN incpp_cat='   0- 1000 ';
else IF incomeperperson LT 2000 THEN incpp_cat='1000-2000 ';
else IF incomeperperson LT 3000 THEN incpp_cat='2000-3000 ';
else IF incomeperperson LT 4000 THEN incpp_cat='3000-4000 ';
else IF incomeperperson LT 5000 THEN incpp_cat='4000-5000 ';
else IF incomeperperson LT 6000 THEN incpp_cat='5000-6000 ';
else IF incomeperperson LT 7000 THEN incpp_cat='6000-7000 ';
else IF incomeperperson LT 8000 THEN incpp_cat='7000-8000 ';
else IF incomeperperson LT 9000 THEN incpp_cat='8000-9000 ';
else IF incomeperperson LT 10000 THEN incpp_cat='9000-10000';
else IF incomeperperson GE 10000 THEN incpp_cat='10000+';

/*making categories based on employment for a country*/
/*IF employrate = . THEN emprate_cat="Data NA";
else*/ IF employrate LT 20 THEN emprate_cat='00-20';
else IF employrate LT 30 THEN emprate_cat='20-30';
else IF employrate LT 40 THEN emprate_cat='30-40';
else IF employrate LT 50 THEN emprate_cat='40-50';
else IF employrate LT 60 THEN emprate_cat='50-60';
else IF employrate LT 70 THEN emprate_cat='60-70';
else IF employrate LT 80 THEN emprate_cat='70-80';
else IF employrate GE 80 THEN emprate_cat='80+';

/*making categories based on urban population*/
/*IF urbanrate = NA THEN incpp_cat='NA';
else*/ IF urbanrate LT 50 THEN urate_cat='Less than 50%';
else IF urbanrate GE 50 THEN urate_cat='More than 50%';

/*making categories based on life expectancy at birth for a country*/
IF lifeexpectancy = . THEN emprate_cat='NA   ';
else IF lifeexpectancy LE 40 THEN lexp_cat='00-40';
else IF lifeexpectancy LE 50 THEN lexp_cat='40-50';
else IF lifeexpectancy LE 60 THEN lexp_cat='50-60';
else IF lifeexpectancy LE 70 THEN lexp_cat='60-70';
else IF lifeexpectancy LE 80 THEN lexp_cat='70-80';
else IF lifeexpectancy LE 90 THEN lexp_cat='80-90';
else IF lifeexpectancy LE 100 THEN lexp_cat='90-100';
else IF lifeexpectancy GT 100 THEN lexp_cat='100+';



Proc SORT; by country;

PROC PRINT; VAR COUNTRY incomeperperson employrate urbanrate lifeexpectancy;

PROC UNIVARIATE; VAR incomeperperson employrate urbanrate lifeexpectancy;

PROC GCHART; VBAR incpp_cat/discrete type=PCT width=8;
PROC GCHART; VBAR emprate_cat/discrete type=PCT width=8;
PROC GCHART; VBAR urate_cat/discrete type=PCT width=8;
PROC GCHART; VBAR lexp_cat/discrete type=PCT width=8;

PROC GPLOT; PLOT urbanrate*lifeexpectancy;
PROC GPLOT; PLOT incomeperperson*lifeexpectancy;
/*PROC GPLOT; PLOT urbanrate*employrate;
PROC GPLOT; PLOT employrate*urbanrate;*///*both of this is better understood by bar chart*/
PROC GPLOT; PLOT incomeperperson*urbanrate;

PROC GCHART; VBAR lexp_cat/discrete type=mean SUMVAR=urbanrate;
PROC GCHART; VBAR lexp_cat/discrete type=mean SUMVAR=incomeperperson;
PROC GCHART; VBAR emprate_cat/discrete type=mean SUMVAR=urbanrate;
PROC GCHART; VBAR urate_cat/discrete type=mean SUMVAR=employrate;
PROC GCHART; VBAR urate_cat/discrete type=mean SUMVAR=incomeperperson;


PROC FREQ; TABLES lexp_cat incpp_cat emprate_cat urate_cat
                  lifeexpectancy incomeperperson employrate urbanrate;

RUN;
```

Univariate Procedure - Incomeperperson

The UNIVARIATE Procedure
Variable: incomeperperson (Gross Domestic Product per capita in constant 2000 US dollar (2010))

Moments			
N	190	Sum Weights	190
Mean	8740.96608	Sum Observations	1660783.55
Std Deviation	14262.8091	Variance	203427723
Skewness	3.25047792	Kurtosis	14.6656757
Uncorrected SS	5.29647E10	Corrected SS	3.84478E10
Coeff Variation	163.171999	Std Error Mean	1034.73292

Basic Statistical Measures			
Location		Variability	
Mean	8740.966	Std Deviation	14263
Median	2553.496	Variance	203427723
Mode	.	Range	105044
		Interquartile Range	8681

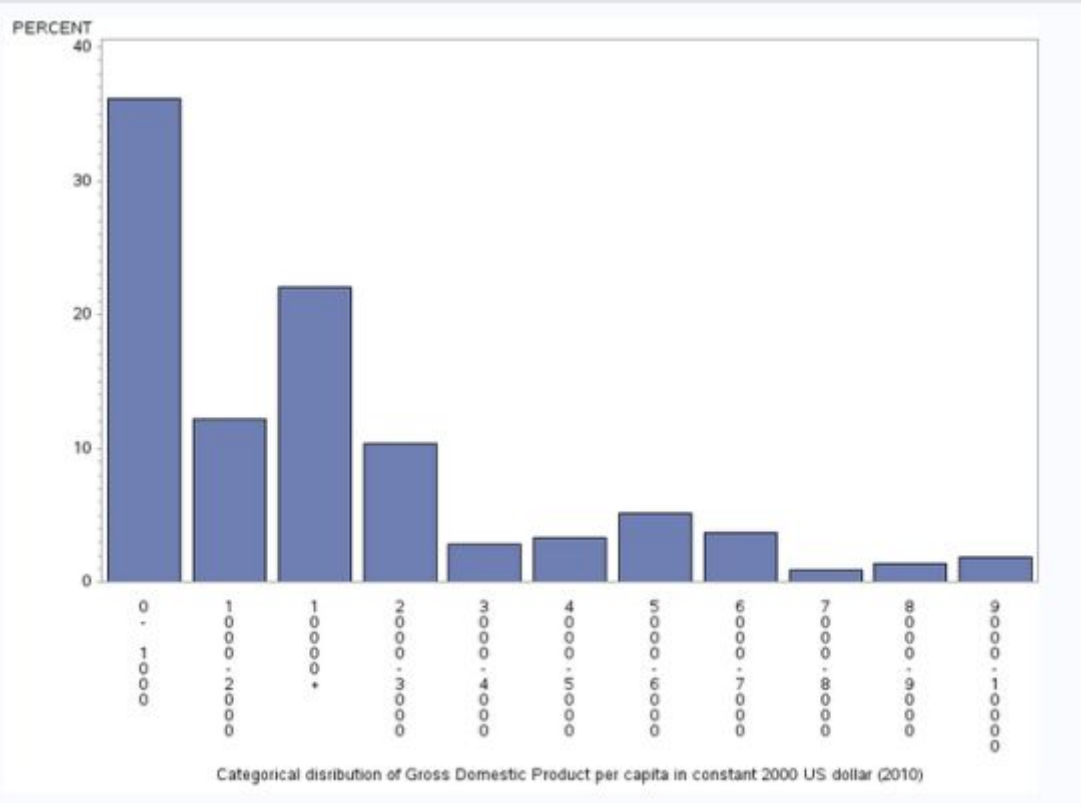
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	8.447558	Pr > |t|	<.0001
Sign	M	95	Pr >= |M|	<.0001
Signed Rank	S	9072.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	105147.438
99%	81647.100
95%	33945.314
90%	26901.858
75% Q3	9425.326
50% Median	2553.496
25% Q1	744.239
10%	337.318
5%	242.678
1%	115.306
0% Min	103.776

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
103.776	42	39972.4	145
115.306	30	52301.6	112
131.796	59	62682.1	21
155.033	108	81647.1	110
161.317	80	105147.4	128

Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	23	10.80	100.00

Univariate Graph - Income per person Categorical (incpp_cat)



This graph is skewed towards the right as it has higher frequency in the lower categories than in the higher categories. This is following almost like decreasing trend while going towards higher categories with some small peaks occurring in between. Also this graph is missing 23 values among the 213 values of the countries.

Univariate Procedure - EmployRate

The UNIVARIATE Procedure
Variable: employrate (% of total population, age above 15, that has been employed during 2007)

Moments			
N	178	Sum Weights	178
Mean	58.6359551	Sum Observations	10437.2
Std Deviation	10.5194545	Variance	110.658922
Skewness	0.13984206	Kurtosis	-0.1660433
Uncorrected SS	631581.82	Corrected SS	19586.6292
Coeff Variation	17.9402799	Std Error Mean	0.78846645

Basic Statistical Measures			
Location		Variability	
Mean	58.63596	Std Deviation	10.51945
Median	58.70000	Variance	110.65892
Mode	47.30000	Range	51.20000
		Interquartile Range	13.80000

Note: The mode displayed is the smallest of 7 modes with a count of 3.

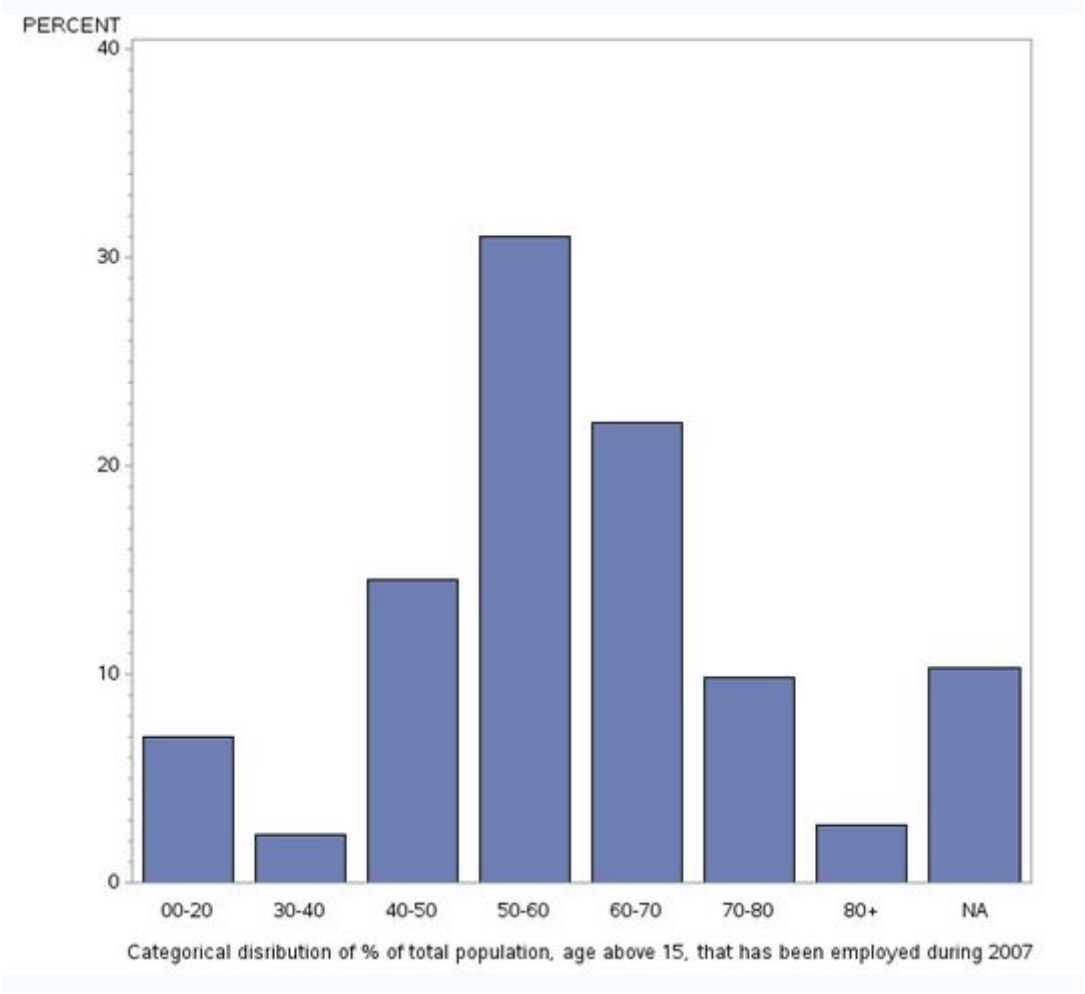
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	74.36709	Pr > |t|	<.0001
Sign	M	89	Pr >= |M|	<.0001
Signed Rank	S	7965.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	83.2
99%	83.2
95%	78.2
90%	73.1
75% Q3	65.0
50% Median	58.7
25% Q1	51.2
10%	44.7
5%	41.6
1%	34.9
0% Min	32.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
32.0	210	81.3	29
34.9	114	81.5	79
37.4	90	83.0	115
38.9	96	83.2	30
39.0	211	83.2	200

Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	35	16.43	100.00

Univariate Graph - EmployRate Categorical (emprate_cat)



This graph seems to be almost centro symmetric with the highest frequency at around the median values. This is following trend somewhat like that of a gaussian curve like. Also this graph is missing 35 values among the 213 values of the countries which are here represented by NA i.e. Not Available.

Univariate Procedure - UrbanRate

The UNIVARIATE Procedure
Variable: urbanrate (Uban population (% of total population) (2008))

Moments			
N	203	Sum Weights	203
Mean	56.7693596	Sum Observations	11524.18
Std Deviation	23.8449326	Variance	568.580813
Skewness	-0.0188477	Kurtosis	-0.9952228
Uncorrected SS	769073.643	Corrected SS	114853.324
Coeff Variation	42.0031736	Std Error Mean	1.67358618

Basic Statistical Measures			
Location		Variability	
Mean	56.7694	Std Deviation	23.84493
Median	57.9400	Variance	568.58081
Mode	100.0000	Range	89.60000
		Interquartile Range	37.68000

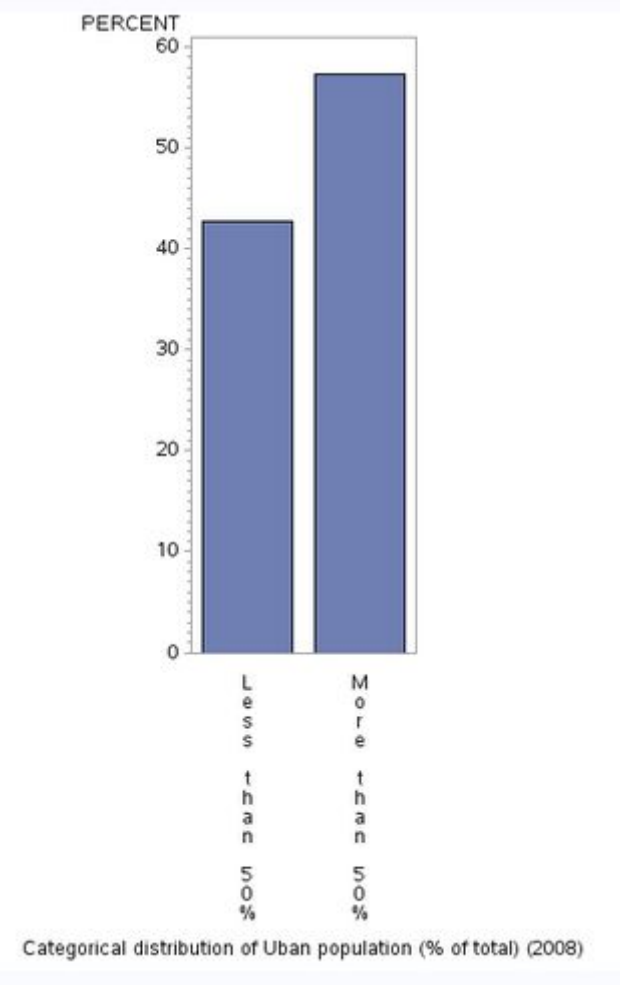
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	33.92079	Pr > |t|	<.0001
Sign	M	101.5	Pr >= |M|	<.0001
Signed Rank	S	10353	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	100.00
99%	100.00
95%	94.26
90%	88.92
75% Q3	74.50
50% Median	57.94
25% Q1	36.82
10%	24.78
5%	18.34
1%	12.98
0% Min	10.40

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
10.40	30	100	35
12.54	150	100	84
12.98	200	100	113
13.22	195	100	128
14.32	110	100	174

Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	10	4.69	100.00

Univariate Graph - UrbanRate Categorical (urate_cat)



With only two category defined for this, and observing the graph it can be said that this plot if skewed towards lest that is it has higher frequency in the ‘more than 50%’ portion. This signify that more no. of countries has more of its population living in urban area.

Univariate Procedure - LifeExpectancy

The UNIVARIATE Procedure
Variable: lifeexpectancy (Life expectancy at birth (years) (2011))

Moments			
N	191	Sum Weights	191
Mean	69.7535236	Sum Observations	13322.923
Std Deviation	9.70862054	Variance	94.2573127
Skewness	-0.8224064	Kurtosis	-0.4194135
Uncorrected SS	947229.713	Corrected SS	17908.8894
Coeff Variation	13.9184661	Std Error Mean	0.70249113

Basic Statistical Measures			
Location		Variability	
Mean	69.75352	Std Deviation	9.70862
Median	73.13100	Variance	94.25731
Mode	72.97400	Range	35.60000
		Interquartile Range	12.41200

Note: The mode displayed is the smallest of 2 modes with a count of 2.

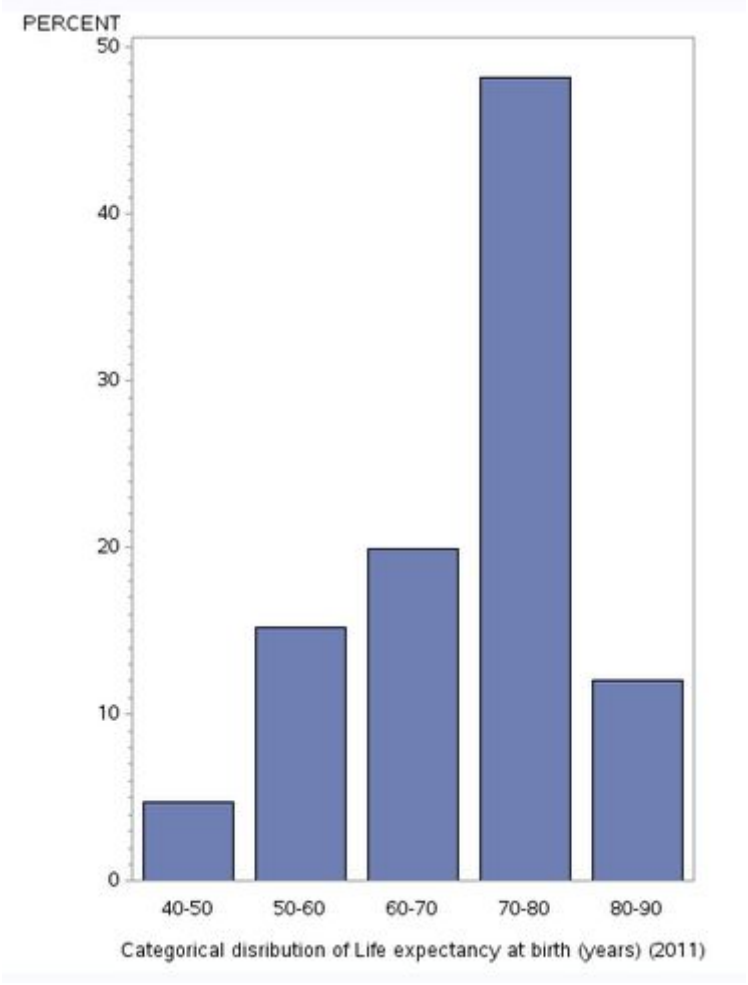
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	99.29453	Pr > |t|	<.0001
Sign	M	95.5	Pr >= |M|	<.0001
Signed Rank	S	9168	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	83.394
99%	82.759
95%	81.404
90%	80.499
75% Q3	76.640
50% Median	73.131
25% Q1	64.228
10%	53.183
5%	50.239
1%	48.132
0% Min	47.794

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
47.794	173	81.855	93
48.132	80	81.907	10
48.196	107	82.338	186
48.397	42	82.759	84
48.398	36	83.394	95

Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	22	10.33	100.00

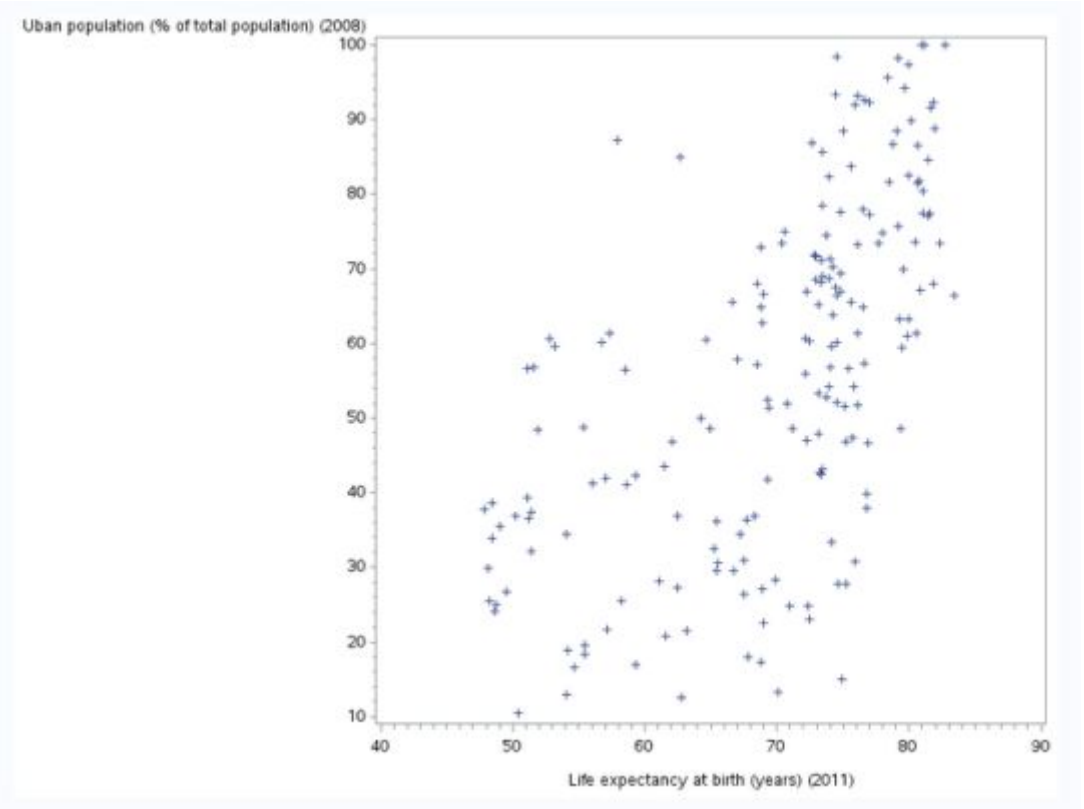
Univariate Graph -LifeExpectancy Categorical (lexp_cat)



This graph is skewed towards the left as it has higher frequency in the higher categories than in the lower categories. This is following almost like an increasing trend while going towards higher categories to a limit then it falls down. Also this graph is missing 22 values among the 213 values of the countries. This is a unimodal graph which symbolizes that the large mo. of countries have life expectancy at birth of around 70 to 80 years.

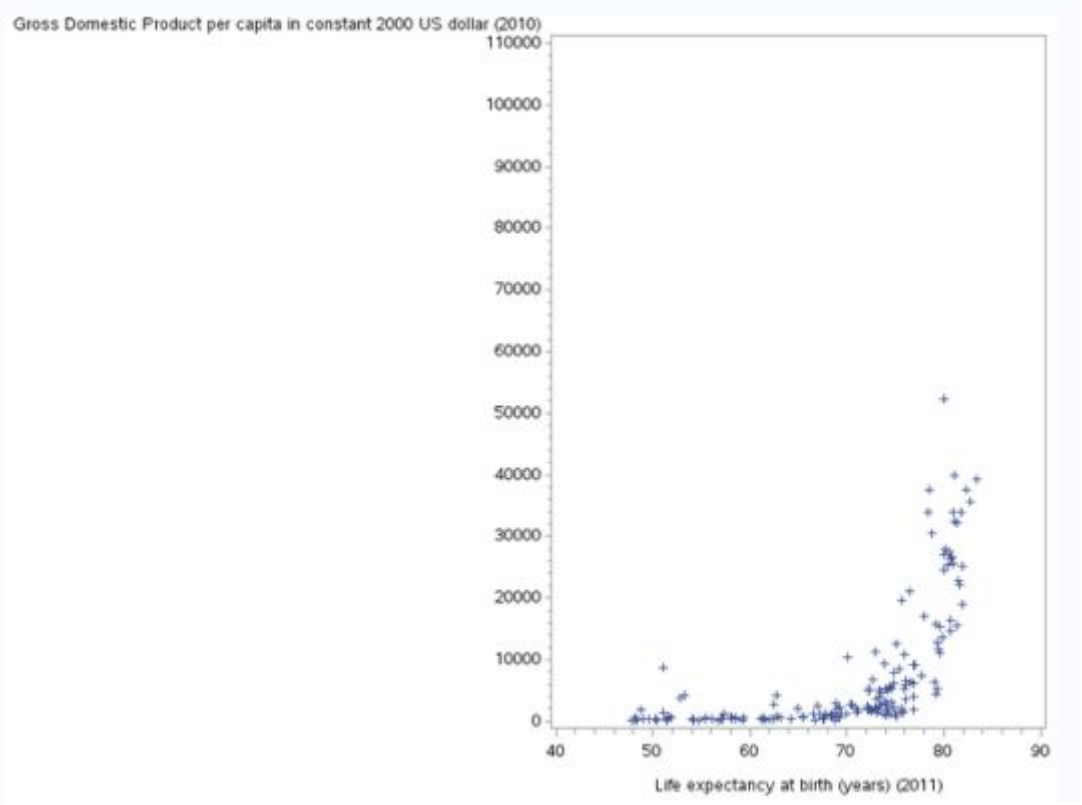
GRAPHS

UrbanRate vs LifeExpectancy



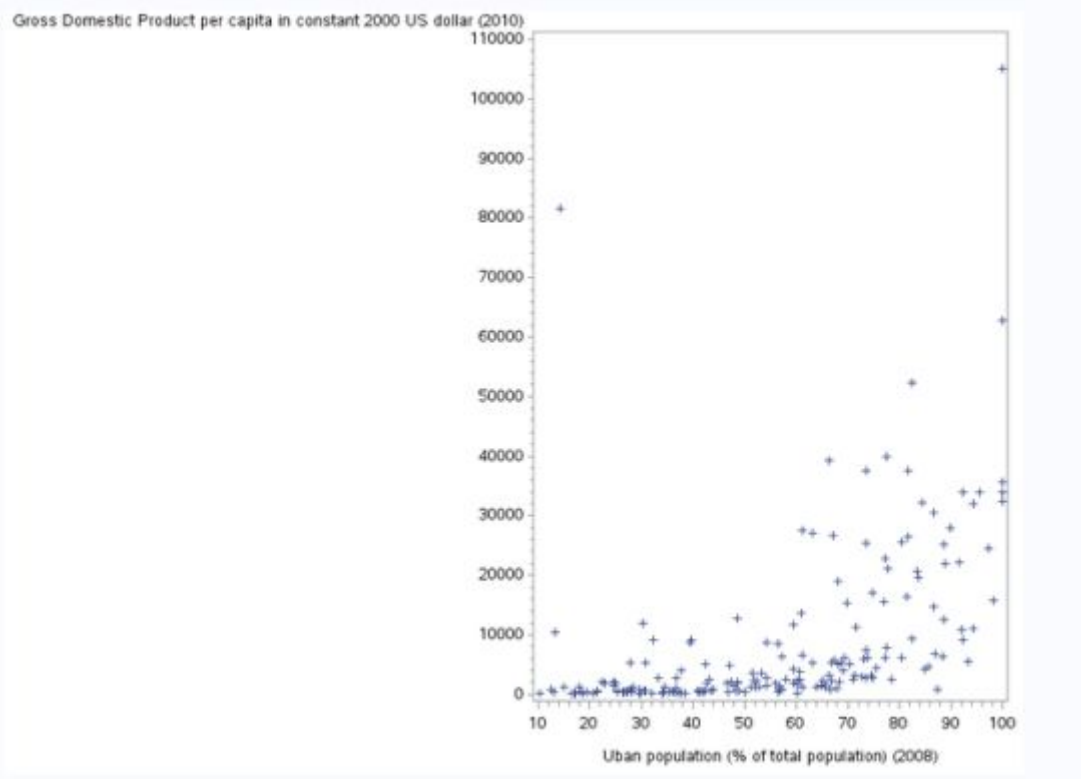
From this scatter plot we can see that countries that are having higher urban population % has higher life expectancy at birth i.e. the chances of living a long life is more in a country where more population lives in urban areas,

IncomePerPerson vs LifeExpectancy



From this scatter plot we can see that countries that are having higher income per person are having higher life expectancy at birth i.e. the chances of living a long life is more in a country Which has higher income per person, Also life expectancy shows large variability for the countries having low income per person.

IncomePerPerson vs UrbanRate



From this scatter plot we can see that countries that are having higher income per person are having higher percentage of urban population i.e. more percentage of people are living in urban areas of the countries which have higher income per person, Also life urban rate shows large variability for the countries having low income per person. Also there are some good exception present in this.