Assignment 3

Remark: While working I wrote down enough code and for this assignment I haven't shorten it and though this program will give various result, I will only be showing the asked results for this particular assignment.

```
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;
```

Note: As I copied the image from PDF its color scheme has changed and also for some variable I won't be able to showcase their full freq.
Procedure.

The FREQ Procedure

Categorical disribution of Life expectancy at birth (years) (2011)				
lexp_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
40-50	9	4.71	9	4.71
50-60	29	15.18	38	19.90
60-70	38	19.90	76	39.79
70-80	92	48.17	168	87.96
80-90	23	12.04	191	100.00
Frequency Missing = 22				

Categorical disribution of Gross Domestic Product per capita in constant 2000 US dollar (2010)				
incpp_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0- 1000	77	36.15	77	36.15
1000-2000	26	12.21	103	48.36
10000+	47	22.07	150	70.42
2000-3000	22	10.33	172	80.75
3000-4000	6	2.82	178	83.57
4000-5000	7	3.29	185	86.85
5000-6000	11	5.16	196	92.02
6000-7000	8	3.76	204	95.77
7000-8000	2	0.94	206	96.71
8000-9000	3	1.41	209	98.12
9000-10000	4	1.88	213	100.00

Categorical disribution of % of total population, age above 15, that has been employed during 2007				
emprate_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
00-20	15	7.04	15	7.04
30-40	5	2.35	20	9.39
40-50	31	14.55	51	23.94
50-60	66	30.99	117	54.93
60-70	47	22.07	164	77.00
70-80	21	9.86	185	86.85
80+	6	2.82	191	89.67
NA	22	10.33	213	100.00

Categorical distribution of Uban population (% of total) (2008)				
urate_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Less than 50%	91	42.72	91	42.72
More than 50%	122	57.28	213	100.00

The FREQ Procedure

Life expectancy at birth (years) (2011)				
lifeexpectancy	Frequency	Percent	Cumulative Frequency	Cumulative Percent
47.794	1	0.52	1	0.52
48.132	1	0.52	2	1.05
48.196	1	0.52	3	1.57
48.397	1	0.52	4	2.09
48.398	1	0.52	5	2.62
48.673	1	0.52	6	3.14
48.718	1	0.52	7	3.66
49.025	1	0.52	8	4.19
49.553	1	0.52	9	4.71
50.239	1	0.52	10	5.24
50.411	1	0.52	11	5.76
51.088	1	0.52	12	6.28
51.093	1	0.52	13	6.81
51.219	1	0.52	14	7.33
51.384	1	0.52	15	7.85
51.444	1	0.52	16	8.38
51.61	1	0.52	17	8.90
51.879	1	0.52	18	9.42
52.797	1	0.52	19	9.95
53.183	1	0.52	20	10.47
54.097	1	0.52	21	10.99
54.116	1	0.52	22	11.52
54.21	1	0.52	23	12.04
54.675	1	0.52	24	12.57

The FREQ Procedure

Gross Domestic Product per capita in constant 2000 US dollar (2010)				
incomeperperson	Frequency	Percent	Cumulative Frequency	Cumulative Percent
103.77585724	1	0.53	1	0.53
115.3059959	1	0.53	2	1.05
131.79620701	1	0.53	3	1.58
155.03323123	1	0.53	4	2.11
161.3171371	1	0.53	5	2.63
180.083376	1	0.53	6	3.16
184.14179659	1	0.53	7	3.68
220.89124792	1	0.53	8	4.21
239.51874937	1	0.53	9	4.74
242.67753416	1	0.53	10	5.26
268.25944951	1	0.53	11	5.79
268.3317903	1	0.53	12	6.32
269.89288112	1	0.53	13	6.84
275.88428653	1	0.53	14	7.37
276.20041296	1	0.53	15	7.89
279.18045256	1	0.53	16	8.42
285.22444925	1	0.53	17	8.95
320.77188995	1	0.53	18	9.47
336.36874948	1	0.53	19	10.00
338.26639123	1	0.53	20	10.53
354.59972629	1	0.53	21	11.05
358.9795398	1	0.53	22	11.58
369.57295374	1	0.53	23	12.11
371.42419752	1	0.53	24	12.63
372.728414	1	0.53	25	13.16
377.03969946	1	0.53	26	13.68
377.42111326	1	0.53	27	14.21
389.76363425	1	0.53	28	14.74

The FREQ Procedure

% of total population, age above 15, that has been employed during 2007				
employrate	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	0.56	1	0.56
34.900001526	1	0.56	2	1.12
37.400001526	1	0.56	3	1.69
38.900001526	1	0.56	4	2.25
39	1	0.56	5	2.81
40.099998474	1	0.56	6	3.37
41.099998474	1	0.56	7	3.93
41.200000763	1	0.56	8	4.49
41.599998474	1	0.56	9	5.06
42	1	0.56	10	5.62
42.400001526	2	1.12	12	6.74
42.5	1	0.56	13	7.30
42.799999237	1	0.56	14	7.87
43.099998474	1	0.56	15	8.43
44.200000763	1	0.56	16	8.99
44.299999237	1	0.56	17	9.55
44.700000763	1	0.56	18	10.11
44.799999237	1	0.56	19	10.67
45.700000763	1	0.56	20	11.24
46	2	1.12	22	12.36
46.200000763	1	0.56	23	12.92
46.400001526	1	0.56	24	13.48
46.799999237	1	0.56	25	14.04
46.900001526	1	0.56	26	14.61
47.099998474	1	0.56	27	15.17
47.299999237	3	1.69	30	16.85
47.799999237	1	0.56	31	17.42
48.599998474	2	1.12	33	18.54
48.700000763	2	1.12	35	19.66
49.5	1	0.56	36	20.22
49.599998474	1	0.56	37	20.79

The FREQ Procedure

Uban population (% of total population) (2008)				
urbanrate	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10.4	1	0.49	1	0.49
12.54	1	0.49	2	0.99
12.98	1	0.49	3	1.48
13.22	1	0.49	4	1.97
14.32	1	0.49	5	2.46
15.1	1	0.49	6	2.96
16.54	1	0.49	7	3.45
17	1	0.49	8	3.94
17.24	1	0.49	9	4.43
17.96	1	0.49	10	4.93
18.34	1	0.49	11	5.42
18.8	1	0.49	12	5.91
19.56	1	0.49	13	6.40
20.72	1	0.49	14	6.90
21.56	1	0.49	15	7.39
21.6	1	0.49	16	7.88
22.54	1	0.49	17	8.37
23	1	0.49	18	8.87
24.04	1	0.49	19	9.36
24.76	1	0.49	20	9.85
24.78	1	0.49	21	10.34
24.94	1	0.49	22	10.84
25.46	1	0.49	23	11.33
25.52	1	0.49	24	11.82
26.46	1	0.49	25	12.32
26.68	1	0.49	26	12.81
27.14	1	0.49	27	13.30
27.3	1	0.49	28	13.79

Data management includes such things as coding out missing data, coding in valid data, recoding variables, creating secondary variables and binning or grouping variables. Not everyone does all of these, but some is required

As for the data management objective I collapsed the responses for **incomeperperson, employrate, urbanrate,** and **lifeexpectancy** to create four **new secondary categorical variables** for each of them: **incpp_cat**, **emprate_cat**, **urate_cat**, and **lexp_cat**. These variable categorize the dataset for the variable into categories. Also the missing data has been coded out wherever its present. Also for more better interpretation I am adding some plotted charts at the end.

For **lexp_cat**, the most commonly endorsed response was '*70-80*' (48.17%), meaning that most countries have an life expectancy in the age group of 70 and 80 years. Also among the 213 countries we are missing the data for 22 of them for this variable.
For **incpp_cat**, the most commonly endorsed response was '*0-1000*' (36.15%), meaning that most countries GDP per capital lies in constant 2000 $ lies in 0 - 1000 $ group. Also looking at the data it can be said that the plot will have a skewed right distribution.
For **emprate_cat**, the most common endorsed response was '*50-60*' (30.99%), meaning that for many countries comes in the range of having in between 50 to 60 % of total population, age above 15, that has been employed
during 2007. By looking at the data it seem this variable is having an approximate Gaussian curve with about center symmetric around 50-60 range. Also from among the 213 countries, dataset is missing the values for 22 countries.
For **urate_cat**, the most commonly endorsed response was '*More than 50%*' (57.28%), meaning that among the 213 countries, 122 countries are having majority of its population living in urban areas and the majority population for the rest 91 countries is living in rural areas.
The above discussed variable significantly help in understanding the result and making the result concise that otherwise wold have been difficult to be discussed with the freq procedure for incomeperperson, employrate, urbanrate, and lifeexpectancy.