SAS Data set for Q1 to Q6: WORK.Coffee_data

Q1. Using the procedure 'PROC MEANS' to summarize the SAS data set 'WORK.Coffee_data'. The resulting table 'coffee_sum0' should be as below

	Smoker	_FREQ_	mean	max	n	nmiss
1	All	234	2.3299492386	33	197	37
2	no	154	2.2777777778	29	126	28
3	yes	80	2.4225352113	33	71	9

Where the column _FREQ_ is the total number of observations and 'mean', 'max', 'n' and 'nmiss' are the statistics for the variable in analysis 'Cups_Per_Week' within each 'Smoker' group (i.e. no, yes and All). Additionally, please explain those statistics.

Q2. Using the procedure 'PROC MEANS' to summarize two variables 'Cups_Per_Week' and 'Income'. It is required to create the following tables: (1) coffee_sum1 (2) coffee_sum2 and (3) coffee_sum3.

coffee_sum1

Average **Average**

	Gender	Smoker	totcount	Cups_Per_Week	Income
1	F	no	87	1.6901408451	40808.219178
2	F	yes	50	2.2127659574	40043.478261
3	M	no	67	3.0363636364	54631.578947
4	M	yes	30	2.8333333333	44120

coffee_sum2

Non missing # **Missing #**

	Gender	Smoker	totcount	income_avg	count	num_missing
1	F	no	87	40808.219178	73	14
2	F	yes	50	40043.478261	46	4
3	M	no	67	54631.578947	57	10
4	M	yes	30	44120	25	5

coffee_sum3

	Gender	Smoker	totcount	Cups_Per_Week_avg	Cups_Per_Week_max
1	F	no	87	1.6901408451	13
2	F	yes	50	2.2127659574	18
3	M	no	67	3.0363636364	29
4	M	yes	30	2.8333333333	33

Q3. Programing for the following questions (1) Applying 'PROC FORMAT' to define an income buckets (ranking groups) (2) Using the buckets and 'PROC MEANS' to analyze the variables 'married' and 'Cups_Per_Week'. You will obtain the following resulting data set:

coffee_sum1

	Income	totcount	avg_Cups_Per_Week	married_rate	Cups_Per_Week_miss
1	MISSING	33	1.9583333333	0.303030303	9
2	0_20000	47	0.3095238095	0.4468085106	5
3	20000-50000	82	1.1617647059	0.4756097561	14
4	50000-80000	39	2.9428571429	0.5384615385	4
5	80000P	33	7.75	0.5151515152	5

Where the columns 'avg_Cups_Per_Week'and 'married_rate' are the mean values of variables 'Cups_Per_Week' and 'married' respectively. In addition, the last column 'Cups_Per_Week_miss' stands for the number of missing values for the variable 'Cups_Per_Week'. (Note the **'MISSING'** is also an income bucket in this analysis).

Q4. Using the 'PROC FREQ' to summarize frequency of the categorical variables 'gender' and 'employment'. You will generate a listing table (not cross tab, see the output below) in the SAS output window and also create a SAS resulting table 'freq_1' (see below). Note, the missing value of 'gender' or 'employments' are also treated as a class in this analysis

The output of 'PROC FREQ'

The FREQ Procedure

Gender	Employment	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F		16	6.84	16	6.84
F	fulltime	64	27.35	80	34.19
F	parttime	25	10.68	105	44.87
F	student	13	5.56	118	50.43
F	unemployment	19	8.12	137	58.55
M		13	5.56	150	64.10
M	fulltime	43	18.38	193	82.48
M	parttime	19	8.12	212	90.60
M	student	5	2.14	217	92.74
M	unemployment	17	7.26	234	100.00

The Resulting SAS Table 'freq_1'

	Gender	Employment	Frequency Count	Percent of Total Frequency
1	F		16	6.8376068376
2	F	fulltime	64	27.35042735
3	F	parttime	25	10.683760684
4	F	student	13	5.5555555556
5	F	unemployment	19	8.1196581197
6	M		13	5.5555555556
7	M	fulltime	43	18.376068376
8	M	parttime	19	8.1196581197
9	M	student	5	2.1367521368
10	M	unemployment	17	7.264957265

Q5. Summarizing two variables 'Cups_Per_Week' and 'age'. First defining a format to group 'Cups_Per_Week' i.e.

```
proc format;
   VALUE cups
        0 -< 0 = 'None'
        1 -< 2 = 'slight'
        2 -< 3 = 'Medium'
        4 - HIGH = 'Heavy'
        OTHER = 'MISSING'
        ;
run;
```

Then creating a cross tab using the procedure 'PROC FREQ':

The FREQ Procedure

Row Pct	Table of Age by Cups_Per_Week				
	Cups_Per_Week				
Age	MISSING	slight	Medium	Heavy	Total
20	73.17	19.51	2.44	4.88	
30	60.42	14.58	6.25	18.75	
40	41.07	19.64	17.86	21.43	
50	53.33	20.00	8.89	17.78	
60	56.82	31.82	4.55	6.82	
Total	131	49	20	34	234

Where the number in each cell is the 'row percentage' for each age group (hint: using 'NOCOL' 'NOPERCENT' 'NOFREQ' options).

Q5. Study the frequency between 'gender' and 'smoker' stratified by the 'age' group (using 'BY' statement). Display all resulting cross tabs in the order of frequency (see following results). Where the number in each cell is the 'row frequency' and 'row percentage' (hint: using 'NOCOL' 'NOPERCENT' 'NOFREQ' options).

---------- Age=20 ----------

The FREQ Procedure

Table of Gender by Smoker

Gender Smoker

Frequency Row Pct	no	yes	Total
F	12 50.00	12 50.00	24
M	10 58.82	7 41.18	17
Total	22	19	41

---------- Age=30 ----------

The FREQ Procedure

Table of Gender by Smoker

Gender Smoker

Frequency Row Pct	no	yes	Total
F	21 72.41	8 27.59	29
M	10 52.63	9 47.37	19
Total	31	17	48

---------- Age=40 ----------

The FREQ Procedure

Table of Gender by Smoker

Gender Smoker

Frequency Row Pct	no	yes	Total
F	19 59.38	13 40.63	32
M	19 79.17	5 20.83	24
Total	38	18	56

---------- Age=50 ----------

The FREQ Procedure

Table of Gender by Smoker

Gender Smoker

Frequency Row Pct	no	yes	Total
F	21 80.77	5 19.23	26
M	16 84.21	3 15.79	19
Total	37	8	45

---------- Age=60 ----------

The FREQ Procedure

Table of Gender by Smoker

Gender Smoker

Frequency Row Pct	no	yes	Total
F	14 53.85	12 46.15	26
M	12 66.67	6 33.33	18
Total	26	18	44