```
libname a2 '/home/u63416676/BAN110ZBB';

data customer_all;
set a2.customer_all;
file print;
run;

*answer to Q1*;
proc freq data=customer_all;
table y;
title "Q1."
run;

proc freq data=customer_all;
table contact y;
title 'Q2. Examine the variable "contact" and study its dependency with the target variable y';
run;
* There are no invalid values*;


proc freq data=customer_all;
tables contact*y/chisq plots=mosaic;
title 'Q3. Contiengency table Contact by y and mosaic plot';
run;
*According to the plot there is not association between the variables Y and Contact*;

*answer to Q4 - 4.1*;
proc format;
value $education_Check 'primary','secondary','tertiary','unknown'= 'Valid'
' ' = 'Missing'
other = 'Miscoded';
run;
data null;
set a2.customer_all (keep=customer_id Education);
file print;
if put(Education, $education_Check.) = 'Missing' or put(Education, $education_Check.) = 'Miscoded'
    then put customer_id= Education= "-> Invalid Value";
title1 'Q4. Examine the variable "education"';
title2 'Q4.1 Define a new format, name it education_Check and use it to identify invalid values for the variable education.';
run;

*answer to Q4 - 4.2*;
data a2.customer_all;
set a2.customer_all;
Education = lowcase(Education);
proc print data=a2.customer_all (obs=10) noobs;
title "Q4.2 Use the function lowcase on education column. use the same dataset name for output dataset";
run;
*answer to Q4 - 4.3*;
proc freq data=a2.customer_all;
table Education;
title 'Q4.3 show the simple frequency table after the change';
run;

*answer to Q5 - 5.1*;
proc print data=a2.customer_all;
where marital not in ('single','married','divorced');
var customer_id marital;
title1 'Q5. Examine the variable "marital"';
title2 'Q5.1 Use PROC print with a where statement to check for data errors in the variable marital';
run;
*answer to Q5 - 5.2*;
data a2.customer_all;
set a2.customer_all;
marital = lowcase(marital);

proc print data=a2.customer_all (obs=10) noobs;
var customer_id marital;
title 'Q5.2 Use the function lowcase on the variable marital.';
run;
*answer to Q5 - 5.3*;
proc freq data=a2.customer_all;
table marital;
title 'Q5.3 show the simple frequency table after the change';
run;
```

```
*answer to Q6 - 6.1*;
proc freq data=a2.customer_all;
table JOB;
title1 'Q6. Examine the variable "Job"';
title2 'Q6.1 Use PROC FREQ to list a simple frequency table.';
run;
*answer to Q6 - 6.2*;
data a2.customer_all;
set a2.customer_all;
if job = "ADMINISTRATION" or job = "admin." then job = "admin";
run;
proc print data=a2.customer_all (obs=15) noobs;
var customer_id job;
title 'Q6.2 write a code to combine the categories "admin." and "ADMINISTRATION" for the job variable as "admin"';
run;

*answer to Q6 - 6.3*;
proc freq data=a2.customer_all;
table JOB;
title 'Q6.3 show the simple frequency table after the change.';
run;

*answer to Q7*;
proc format;
value $Missing_Count
' ' = 'Missing'
other = 'Not Missing';
run;
proc freq data=a2.customer_all;
tables _character_ / nocum missing;
format _character_ $Missing_Count.;
title 'Q7. checking missing values';
run;



proc freq data = a2.customer_all order=freq;
table jobmf;
title 'Q8. create a new variable named jobMF to indicate the most frequent job category';
run;
data a2.customer_all;
set a2.customer_all;
if JOB = 'management' then jobMF = 1;
else jobMF = 0;
run;
proc print data = a2.customer_all (obs = 10);
var customer_id JOB jobMF;
run;


data Units;
    input Length $ 10. ;
    digits = compress(Length,,'kd');
    if findc(Length,'m','i') then
        Length_m=input(digits,5.);
    else if not missing(digits) then
        Length_m=input(digits,5.)*0.3048;
datalines;
100m.
110 ft.
50M.
70 Ft
180
;
run;
proc print data=Units;
title 'Q9. Removing units from a value and standarizing';
run;
```

Q1. run

The FREQ Procedure

y				
y	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	5289	50.00	5289	50.00
yes	5289	50.00	10578	100.00

Q2. Examine the variable "contact" and study its dependency with the target variable y

The FREQ Procedure

contact				
contact	Frequency	Percent	Cumulative Frequency	Cumulative Percent
cellular	7682	72.62	7682	72.62
telephone	712	6.73	8394	79.35
unknown	2184	20.65	10578	100.00

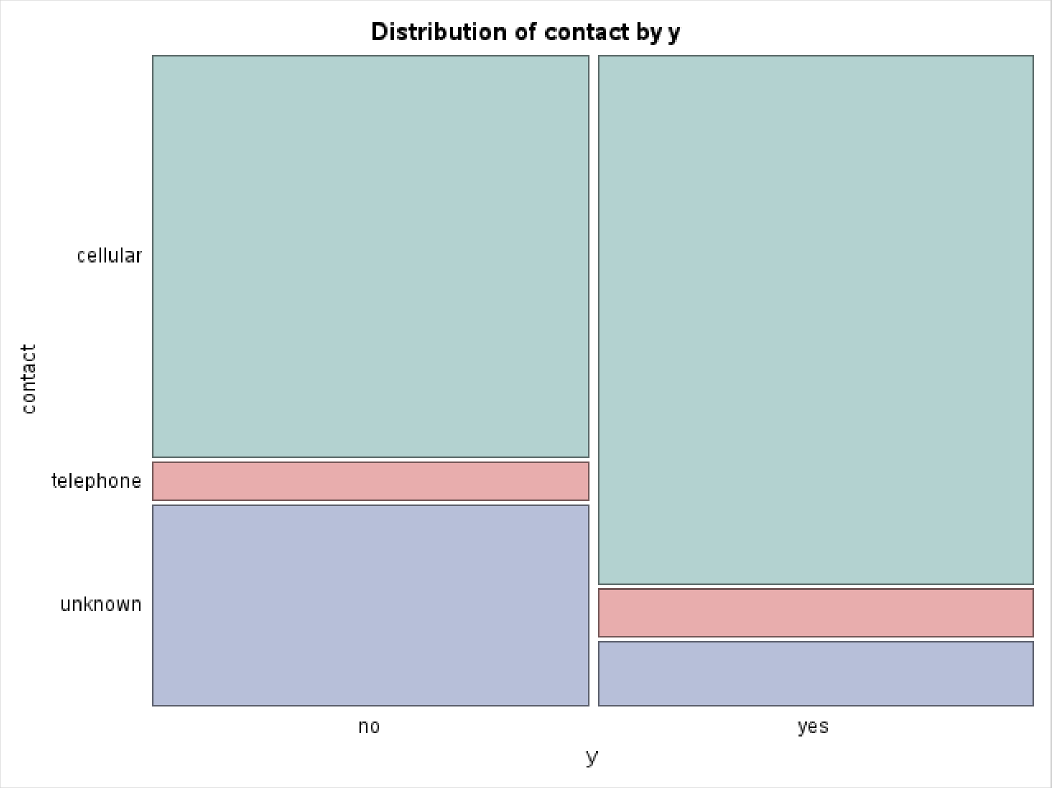
y				
y	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	5289	50.00	5289	50.00
yes	5289	50.00	10578	100.00

Q3. Contiengency table Contact by y and mosaic plot

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of contact by y			
	y(y)		
contact(contact)	no	yes	Total
cellular	3313 31.32 43.13 62.64	4369 41.30 56.87 82.61	7682 72.62
telephone	322 3.04 45.22 6.09	390 3.69 54.78 7.37	712 6.73
unknown	1654 15.64 75.73 31.27	530 5.01 24.27 10.02	2184 20.65
Total	5289 50.00	5289 50.00	10578 100.00



Statistics for Table of contact by y

Statistic	DF	Value	Prob
Chi-Square	2	730.1254	<.0001
Likelihood Ratio Chi-Square	2	759.2990	<.0001
Mantel-Haenszel Chi-Square	1	678.0393	<.0001
Phi Coefficient		0.2627	
Contingency Coefficient		0.2541	
Cramer's V		0.2627	

Sample Size = 10578

Q4.2 Use the function lowcase on education column. use the same dataset name for output dataset

customer_id	contact	day	month	campaign	pdays	previous	poutcome	y	default	balance	housing	loan	Education	AGE	marital	JOB	jobMF
100103	unknown	5	may	1	-1	0	unknown	no	no	2	yes	yes	secondary	33	married	entrepreneur	0
100106	unknown	5	may	1	-1	0	unknown	no	no	231	yes	no	tertiary	35	married	management	1
100118	unknown	5	may	1	-1	0	unknown	no	no	52	yes	no	primary	57	married	blue-collar	0
100119	unknown	5	may	1	-1	0	unknown	no	no	60	yes	no	primary	60	married	retired	0
100121	unknown	5	may	1	-1	0	unknown	no	no	723	yes	yes	secondary	28	married	blue-collar	0
100126	unknown	5	may	1	-1	0	unknown	no	no	-372	yes	no	secondary	44	married	admin	0
100130	unknown	5	may	1	-1	0	unknown	no	no	265	yes	yes	secondary	36	single	technician	0
100141	unknown	5	may	1	-1	0	unknown	no	no	2586	yes	no	secondary	44	divorced	services	0
100161	unknown	5	may	1	-1	0	unknown	no	no	0	yes	no	tertiary	32	married	admin	0
100168	unknown	5	may	1	-1	0	unknown	no	no	59	yes	no	tertiary	59	divorced	management	1

Q4.3 show the simple frequency table after the change

The FREQ Procedure

Education				
Education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
primary	1440	13.61	1440	13.61
secondary	5204	49.20	6644	62.81
tertiary	3470	32.80	10114	95.61

Education				
Education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
unknown	464	4.39	10578	100.00

Q5.2 Use the function lowcase on the variable marital.

customer_id	marital
100103	married
100106	married
100118	married
100119	married
100121	married
100126	married
100130	single
100141	divorced
100161	married
100168	divorced

Q5.3 show the simple frequency table after the change

The FREQ Procedure

marital				
marital	Frequency	Percent	Cumulative Frequency	Cumulative Percent
divorced	1243	11.75	1243	11.75
married	5942	56.17	7185	67.92
single	3393	32.08	10578	100.00

Q6. Examine the variable "Job"
Q6.1 Use PROC FREQ to list a simple frequency table.

The FREQ Procedure

JOB				
JOB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
admin	1185	11.20	1185	11.20
blue-collar	1914	18.09	3099	29.30
entrepreneur	291	2.75	3390	32.05
housemaid	262	2.48	3652	34.52
management	2391	22.60	6043	57.13
retired	757	7.16	6800	64.28
self-employed	367	3.47	7167	67.75
services	850	8.04	8017	75.79
student	375	3.55	8392	79.33
technician	1768	16.71	10160	96.05
unemployed	353	3.34	10513	99.39
unknown	65	0.61	10578	100.00

Q6.2 write a code to combine the categories "admin." and "ADMINISTRATION" for the job variable as "admin"

customer_id	JOB
100103	entrepreneur
100106	management
100118	blue-collar
100119	retired
100121	blue-collar

customer_id	JOB
100126	admin
100130	technician
100141	services
100161	admin
100168	management
100172	services
100184	admin
100187	admin
100188	technician
100189	management

Q6.3 show the simple frequency table after the change.

The FREQ Procedure

JOB				
JOB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
admin	1185	11.20	1185	11.20
blue-collar	1914	18.09	3099	29.30
entrepreneur	291	2.75	3390	32.05
housemaid	262	2.48	3652	34.52
management	2391	22.60	6043	57.13
retired	757	7.16	6800	64.28
self-employed	367	3.47	7167	67.75
services	850	8.04	8017	75.79
student	375	3.55	8392	79.33
technician	1768	16.71	10160	96.05
unemployed	353	3.34	10513	99.39
unknown	65	0.61	10578	100.00

Q7. checking missing values

The FREQ Procedure

contact		
contact	Frequency	Percent
Not Missing	10578	100.00

month		
month	Frequency	Percent
Not Missing	10578	100.00

poutcome		
poutcome	Frequency	Percent
Not Missing	10578	100.00

y		
y	Frequency	Percent
Not Missing	10578	100.00

default	Frequency	Percent
Not Missing	10578	100.00

housing	Frequency	Percent
Not Missing	10578	100.00

loan	Frequency	Percent
Not Missing	10578	100.00

Education		
Education	Frequency	Percent
Not Missing	10578	100.00

marital		
marital	Frequency	Percent
Not Missing	10578	100.00

JOB		
JOB	Frequency	Percent
Not Missing	10578	100.00

Q8. create a new variable named jobMF to indicate the most frequent job category

The FREQ Procedure

jobMF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	8187	77.40	8187	77.40
1	2391	22.60	10578	100.00

Q8. create a new variable named jobMF to indicate the most frequent job category

Obs	customer_id	JOB	jobMF
1	100103	entrepreneur	0
2	100106	management	1
3	100118	blue-collar	0
4	100119	retired	0
5	100121	blue-collar	0
6	100126	admin	0
7	100130	technician	0
8	100141	services	0
9	100161	admin	0
10	100168	management	1

Q9. Removing units from a value and standarizing

Obs	Length	digits	Length_m
1	100m.	100	100.000
2	110 ft.	110	33.528
3	50M.	50	50.000
4	70 Ft	70	21.336
5	180	180	54.864