Create a new variable named 'Is_Complete' and assign the value 'Yes' if the row has complete information and 'No' if it doesn't. Hint: Use cmiss(of _ALL_) function.

```
data patients;

infile '/home/u63571133/BAN110/Assignment 2/Patients (4).txt';

INPUT   @1  PATNO     $3.
                @4  Account_No $7.
                @11 GENDER     $1.
                @12 VISIT      mmddyy10.
                @22 HR         3.
                @25 SBP        3.
                @28 DBP        3.
                @31 DX         $7.
                @38 AE         1.;

LABEL PATNO="Patient Number" Account_No = "Account Number" GENDER="Gender" VISIT="Visit Date" HR="Heart Rate" SBP="Systolic Blood Pressure" DBP="Diastolic Blood Pressure" DX="Diagnosis Code" AE="Adverse Event?";

FORMAT VISIT MMDDYY10.;

run;


proc format;

value completefmt

1 = 'Yes'

other = 'No'

;


data patients_copy;

set patients;

Is_Complete = cmiss(of _ALL_);

format is_complete completefmt.;

run;


proc print; run;
```

Obs	PATNO	Account_No	GENDER	VISIT	HR	SBP	DBP	DX	AE	Is_Complete
1	001	CT14882	M	06/12/2012	69	124	86	713.410	0	Yes
2	002	MD78461	M	06/04/2010	76	130	80	047.570	1	Yes
3	003	DE51381	f	06/22/2013	70	56	70	108.510	0	Yes
4	004	CT37146	M	05/18/2013	76	112	84	669.860	0	Yes
5	005	DE00080	F	04/08/2012	91	106	84	078.160	0	Yes
6	005	DE00080	F	04/08/2012	91	106	84	078.160	0	Yes
7	006	DE37709	M	07/27/2014	71	104	88	967.570	0	Yes
8	007	VT56383	F	01/13/2014	63	128	80	640.260	1	Yes
9	008	PA67069	F	09/28/2013	79	124	72	020.120	0	Yes
10	009	MD68313	F	03/15/1956	82	132	64	894.400	0	Yes
11	007	NJ90043	M	08/06/2010	83	130	102	564.870	0	Yes
12	011	NY60612	F	07/03/2010	68	110	78	530.abc	0	Yes
13	012	CT77620	M	08/21/2014	70	124	78	904.490	0	Yes
14	013	MD45188	M	01/21/2012	50	160	92	985.270	0	Yes
15	014	ME78686	F	03/21/2012	83	110	90	064.480	0	Yes
16	050	PA37838	M	03/09/2011	32	118	90	692.470	0	Yes
17	016	DE11937	F	10/21/2020	65	92	84	865.730	1	Yes
18	017	VT16711	F	06/27/2011	83	112	80	533.360	0	Yes
19	018	NY79484	F	09/20/2012	81	124	90	257.940	0	Yes
20	019	NJ49593	M	10/17/2012	62	210	92	177.860	0	Yes
21	XX5	MA93350	F	11/04/2010	69	122	190	052.040	0	Yes
22	021	NY23491	F	01/23/2014	84	120	76	207.420	1	Yes
23	022	MD00664	F	07/24/2013	74	130	90	845.900	0	Yes
24	023	CT48628	F	09/09/2013	.	300	222	333.570	0	No
25	024	MD34807	M	10/12/2011	80	126	80	301.960	0	Yes
26	025	NY81779	M	04/05/2010	69	110	72	949.790	0	Yes
27	026	VT75117	F	05/26/2014	66	116	92	710.740	0	Yes
28	027	MD40964		09/12/2011	56	128	90	092.880	0	No
29	028	MA66833	F	09/03/2014	77	118	70	872.290	0	Yes
30	029	NY29028	F	.	79	148	88	844.790	0	No

1. How many rows have complete information?

```
proc freq data=patients_copy;
table is_complete / nocum nopercent;
title 'Q2: 92 rows have complete information';
run;
```

Q2: 92 rows have complete information

The FREQ Procedure

Is_Complete	Frequency
Yes	92
No	9

2. **Identify columns for dummy/GLM coding and derive new columns by creating indicating variable of each category.**

```
/* Identify suitable columns for GLM encoding via proc freq */
proc freq data=patients_copy;
table gender ae is_complete;
run;

/* GLM Encoding in dataset */
data patients_encoded;
set patients_copy;
Gender = upcase(Gender);
Male = (Gender = 'M');
Female = (Gender = 'F');
AdverseEvent = (AE = 1);
NotAdverseEvent = (AE = 0);
Complete = (Is_complete = 1);
NotComplete = (Is_Complete > 1);
run;

proc print; run;
```

Obs	PATNO	Account_No	GENDER	VISIT	HR	SBP	DBP	DX	AE	Is_Complete	Male	Female	AdverseEvent	NotAdverseEvent	Complete	NotComplete
1	001	CT14882	M	06/12/2012	69	124	86	713.410	0	Yes	1	0	0	1	1	0
2	002	MD78461	M	06/04/2010	76	130	80	047.570	1	Yes	1	0	1	0	1	0
3	003	DE51381	F	06/22/2013	70	56	70	108.510	0	Yes	0	1	0	1	1	0
4	004	CT37146	M	05/18/2013	76	112	84	669.860	0	Yes	1	0	0	1	1	0
5	005	DE00080	F	04/08/2012	91	106	84	078.160	0	Yes	0	1	0	1	1	0
6	005	DE00080	F	04/08/2012	91	106	84	078.160	0	Yes	0	1	0	1	1	0
7	006	DE37709	M	07/27/2014	71	104	88	967.570	0	Yes	1	0	0	1	1	0
8	007	VT56383	F	01/13/2014	63	128	80	640.260	1	Yes	0	1	1	0	1	0
9	008	PA67069	F	09/28/2013	79	124	72	020.120	0	Yes	0	1	0	1	1	0
10	009	MD68313	F	03/15/1956	82	132	64	894.400	0	Yes	0	1	0	1	1	0
11	007	NJ90043	M	08/06/2010	83	130	102	564.870	0	Yes	1	0	0	1	1	0
12	011	NY60612	F	07/03/2010	68	110	78	530.abc	0	Yes	0	1	0	1	1	0
13	012	CT77620	M	08/21/2014	70	124	78	904.490	0	Yes	1	0	0	1	1	0
14	013	MD45188	M	01/21/2012	50	160	92	985.270	0	Yes	1	0	0	1	1	0
15	014	ME78686	F	03/21/2012	83	110	90	064.480	0	Yes	0	1	0	1	1	0
16	050	PA37838	M	03/09/2011	32	118	90	692.470	0	Yes	1	0	0	1	1	0
17	016	DE11937	F	10/21/2020	65	92	84	865.730	1	Yes	0	1	1	0	1	0
18	017	VT16711	F	06/27/2011	83	112	80	533.360	0	Yes	0	1	0	1	1	0
19	018	NY79484	F	09/20/2012	81	124	90	257.940	0	Yes	0	1	0	1	1	0
20	019	NJ49593	M	10/17/2012	62	210	92	177.860	0	Yes	1	0	0	1	1	0
21	XX5	MA93350	F	11/04/2010	69	122	190	052.040	0	Yes	0	1	0	1	1	0
22	021	NY23491	F	01/23/2014	84	120	76	207.420	1	Yes	0	1	1	0	1	0
23	022	MD00664	F	07/24/2013	74	130	90	845.900	0	Yes	0	1	0	1	1	0
24	023	CT48628	F	09/09/2013	.	300	222	333.570	0	No	0	1	0	1	0	1
25	024	MD34807	M	10/12/2011	80	126	80	301.960	0	Yes	1	0	0	1	1	0
26	025	NY81779	M	04/05/2010	69	110	72	949.790	0	Yes	1	0	0	1	1	0
27	026	VT75117	F	05/26/2014	66	116	92	710.740	0	Yes	0	1	0	1	1	0
28	027	MD40964		09/12/2011	56	128	90	092.880	0	No	0	0	0	1	0	1
29	028	MA66833	F	09/03/2014	77	118	70	872.290	0	Yes	0	1	0	1	1	0
30	029	NY29028	F	.	79	148	88	844.790	0	No	0	1	0	1	0	1

3. Implement two methods to detect errors in numeric variables and two methods to detect errors in character variables in the patient dataset, excluding missing values.

```
title "Identifying Numerical errors: Listing of patient numbers and invalid data values within a range";

data error_check_num1;

file print;

set patients_copy(keep=Patno HR SBP DBP);

if (HR lt 40 and not missing(HR)) or HR gt 100 then put 'HR is not within acceptable range of 40 to
100  ' Patno= HR=;

if (SBP lt 80 and not missing(SBP)) or SBP gt 200 then put 'SBP is not within acceptable range of 80 to
200  ' Patno= SBP=;

if (DBP lt 60 and not missing(DBP)) or DBP gt 120 then put 'HR is not within acceptable range of 60 to
120  ' Patno= DBP=;

run;


title "Identifying Numerical Errors: Invalid Values versus Missing Values";

data error_check_num2;

file print;

set patients_copy(keep=Patno AE Visit HR SBP DBP);

if notdigit(strip(Patno)) and not missing (Patno) then put "Invalid value, contains an alphabetical
character in a numerical variable" Patno " for Patno in patient " ;

if notdigit(strip(AE)) and not missing (AE) then put "Invalid value, contains an alphabetical character
in a numerical variable " AE " for AE in patient " Patno ;

if notdigit(strip(HR)) and not missing (HR) then put "Invalid value, contains an alphabetical character
in a numerical variable " HR " for HR in patient " Patno ;

if notdigit(strip(Visit)) and not missing (Visit) then put "Invalid value, contains an alphabetical
character in a numerical variable " Visit " for VisitDate in patient " Patno ;

if notdigit(strip(SBP)) and not missing (SBP) then put "Invalid value, contains an alphabetical
character in a numerical variable " SBP " for SBP in patient " Patno ;

if notdigit(strip(DBP)) and not missing (DBP) then put "Invalid value, contains an alphabetical
character in a numerical variable " DBP " for DBP in patient " Patno ;

run;


title "Identifying character errors that are not in the acceptable character list";

data error_check_char1;

file print;

set patients_copy;
```

```
if Gender not in ('M','F') and not missing(Gender) then put 'There is a error in Gender for ' PATNO=
'The Gender is not M or F, instead the Gender is ' Gender=;

if AE not in ('0','1') and not missing(AE) then put 'There is a error in AE for ' PATNO= 'The Error is: '
AE=;

run;


title "Identifying character errors using Regular Expressions";

data error_check_char2;

file print;

set patients_copy;

if not prxmatch("/\b[A-Z]+\d\d\d\d\d/",Account_No) and not missing(Account_No) then put

"Invalid value, AccountNo does not follow format of 2 alphabetical characters and 5 digits, "
Account_No= "for Patno in patient: " Patno;

run;
```

Identifying Numerical errors: Listing of patient numbers and invalid data values within a range

```
SBP is not within acceptable range of 80 to 200  PATNO=003 SBP=56
HR is not within acceptable range of 40 to 100  PATNO=050 HR=32
SBP is not within acceptable range of 80 to 200  PATNO=019 SBP=210
HR is not within acceptable range of 60 to 120  PATNO=XX5 DBP=190
SBP is not within acceptable range of 80 to 200  PATNO=023 SBP=300
HR is not within acceptable range of 60 to 120  PATNO=023 DBP=222
HR is not within acceptable range of 40 to 100  PATNO=034 HR=115
HR is not within acceptable range of 40 to 100  PATNO=045 HR=900
HR is not within acceptable range of 60 to 120  PATNO=099 DBP=30
```

Identifying Numerical Errors: Invalid Values versus Missing Values

```
Invalid value, contains an alphabetical character in a numerical variable 03/15/1956  for VisitDate in patient 009
Invalid value, contains an alphabetical character in a numerical variableXX5  for Patno in patient
```

Identifying character errors that are not in the acceptable character list

```
There is a error in Gender for PATNO=003 The Gender is not M or F, instead the Gender is GENDER=f
There is a error in AE for PATNO=087 The Error is: AE=9
There is a error in Gender for PATNO=088 The Gender is not M or F, instead the Gender is GENDER=x
There is a error in Gender for PATNO=095 The Gender is not M or F, instead the Gender is GENDER=1
```

Identifying character errors using Regular Expressions

```
Invalid value, AccountNo does not follow format of 2 alphabetical characters and 5 digits, Account_No=1234567 for Patno in patient:
039
Invalid value, AccountNo does not follow format of 2 alphabetical characters and 5 digits, Account_No=xx13243 for Patno in patient:
041
Invalid value, AccountNo does not follow format of 2 alphabetical characters and 5 digits, Account_No=NY1234z for Patno in patient:
058
```