```
/*libname casuser cas;
/*cas mysession sessopts=(caslib=casuser timeout=1800 locale="en_US");*/

 /* Save the new dataset in the casuser library */


 %let use_training_only = 1;

%if &use_training_only = 1 %then %do;
/* In Scoring Code: Simply pass through training output */


data new_dataset;
    set &dm_data;
    if acci_severity not in (1 ,2 ,3) then delete; /* Remove rows where
acci_severity not equal to 1, 2, or 3 'var_name' equals 'value' */

    if IMP_police_force ne 45 then IMP_police_force = 45; /* changes
police force that was imputated to 45 */

    if loc_auth_highw ne "E10000030" then loc_auth_highw= "E10000030" ;/*
replace missing variable with this value (cnstant value)

/* Remove rows where did_poli_att not equal to 1, 2, or 3 'var_name'
equals 'value' */
    if did_poli_offi_att not in (1 ,2 ,3, -1) then delete;

 /* Remove rows where longitude, latitude, loc_east_osgr, or loc_nor_osgr
are missing' */

    if cmiss(latitude,longitude,loc_nor_osgr,loc_east_osgr) > 0 then
delete;

 /* Adds random date to emplty date field with the condition of keeping
the year 2021 */
    if missing(date) then do; /* Replace 'date_var' with your variable
name */

        /* Define start and end dates for 2021 */
        start_date = '01JAN2021'd; /* SAS numeric date value for
01/01/2021 */
        end_date = '31DEC2021'd;   /* SAS numeric date value for
12/31/2021 */

        /* Generate a random date between start_date and end_date */
        random_days = floor((end_date - start_date + 1) *
rand('uniform')); /* Days between 0 and 364 */
        random_date_num = start_date + random_days; /* Numeric SAS date
*/

        /* Convert numeric date to character with mm/dd/yyyy format */
        date = put(random_date_num, mmddyy10.); /* Format as mm/dd/yyyy
*/
    end;
    /* Ensure the variable remains character type */
    format date $10.; /* Explicitly set length to match mmddyy10. format
*/

/* create new day_of_week variable that ensures all the day of week are
accurate and actually represent the corresponding date because some blank
dates have a dummy day of the week */

day_of_week_Amended = weekday(input(date, mmddyy10.)); /* Sunday=1 to
Saturday=7  */


/* dropping variables not needed or redundant */


drop Row random_date_num  random_days  end_date   start_date
day_of_week;

run;

%end;


/* Step 1: Calculate the mode for each column (excluding -1 values) */
proc freq data=new_dataset noprint;
    where carri_haz ne -1;
    tables carri_haz / out=carri_haz_mode;
run;

proc sort data=carri_haz_mode;
    by descending count;
run;

data _null_;
    set carri_haz_mode(obs=1);
    call symputx('mode_carri_haz', carri_haz);
run;

proc freq data=new_dataset noprint;
    where road_surf_con ne -1;
    tables road_surf_con / out=road_surf_con_mode;
run;

proc sort data=road_surf_con_mode;
    by descending count;
run;

data _null_;
    set road_surf_con_mode(obs=1);
    call symputx('mode_road_surf_con', road_surf_con);
run;

proc freq data=new_dataset noprint;
```

```
    where spec_con_site ne -1;
    tables spec_con_site / out=spec_con_site_mode;
run;

proc sort data=spec_con_site_mode;
    by descending count;
run;

data _null_;
    set spec_con_site_mode(obs=1);
    call symputx('mode_spec_con_site', spec_con_site);
run;

/* Step 2: Replace -1 values with the mode */
data new_dataset_fixed;
    set new_dataset;
    if carri_haz = -1 then carri_haz = &mode_carri_haz;
    if road_surf_con = -1 then road_surf_con = &mode_road_surf_con;
    if spec_con_site = -1 then spec_con_site = &mode_spec_con_site;
run;






/*
proc print data=&dm_output_data;
    title "Contents of dm_output";
run;

*/

proc export data = new_dataset_fixed

outfile =
"/export/viya/homes/di00222@surrey.ac.uk/casuser/Road_Accident_Cleaned_Fi
nal_Dataset.csv"
    dbms=csv
    replace;
run;
```

Road_2021_Data_Cleaning

"Imputation" Results

by: di00222@surrey.ac.uk

Contents

Input Variable Statistics

Input Variable	Variable Level	Number of Missing Values	Percent Missing
carri_haz	NOMINAL	0	0
date	NOMINAL	1	0.0403
day_of_week	NOMINAL	1	0.0403
did_poli_offi_att	NOMINAL	0	0
first_road_class	NOMINAL	0	0
first_road_num	INTERVAL	0	0
junc_con	NOMINAL	0	0
junc_detail	NOMINAL	0	0
latitude	INTERVAL	1	0.0403
light_con	NOMINAL	0	0
local_auth_distr	UNARY	0	0
loc_auth_highw	UNARY	1	0.0403
loc_auth_ons_distr	NOMINAL	0	0
loc_east_osgr	INTERVAL	1	0.0403
loc_nor_osgr	INTERVAL	0	0
longitude	INTERVAL	1	0.0403
lsoa_of_acc_loc	NOMINAL	0	0
num_of_casu	NOMINAL	1	0.0403
num_of_vehi	NOMINAL	1	0.0403
ped_cross_hum_con	NOMINAL	0	0
ped_cross_phy_facil	NOMINAL	0	0
police_force	UNARY	1	0.0403
road_surf_con	NOMINAL	0	0
road_type	NOMINAL	0	0
Row	INTERVAL	0	0

Input Variable	Variable Level	Number of Missing Values	Percent Missing
sec_road_class	NOMINAL	0	0
sec_road_num	INTERVAL	0	0
spec_con_site	NOMINAL	0	0
speed_limit	NOMINAL	0	0
time	NOMINAL	0	0
tru_road_flag	BINARY	0	0
urb_or_rur_area	BINARY	0	0
weath_con	NOMINAL	0	0

Imputable	Minimum	Maximum	Mean
0			
0			
0			
0			
0			
0	0	3,411	355.2101
0			
0			
0	51.0832	51.4664	51.3026
0			
0			
0			
0			
0	482,163	543,673	509,570.0940
0	132,324	175,208	157,127.4137
0	-0.8317	0.0571	-0.4297
0			
1			
1			

Imputable	Minimum	Maximum	Mean
0			
0			
1			
0			
0			
0	1	2,480	1,240.5000
0			
0	-1	3,411	84.4302
0			
0			
0			
0			
0			
0			

Midrange	Standard Deviation	Skewness	Kurtosis
1,705.5000	785.1844	2.6758	5.9131
51.2748	0.0820	-0.2297	-0.5645
512,918	14,018.8531	0.2975	-0.6986
153,766	9,087.0382	-0.2496	-0.5216
-0.3873	0.2006	0.2775	-0.7005
1,240.5000	716.0587	0.0000	-1.2000
1,705	423.7698	6.3374	40.6928

Variable Label

Imputed Variables Summary

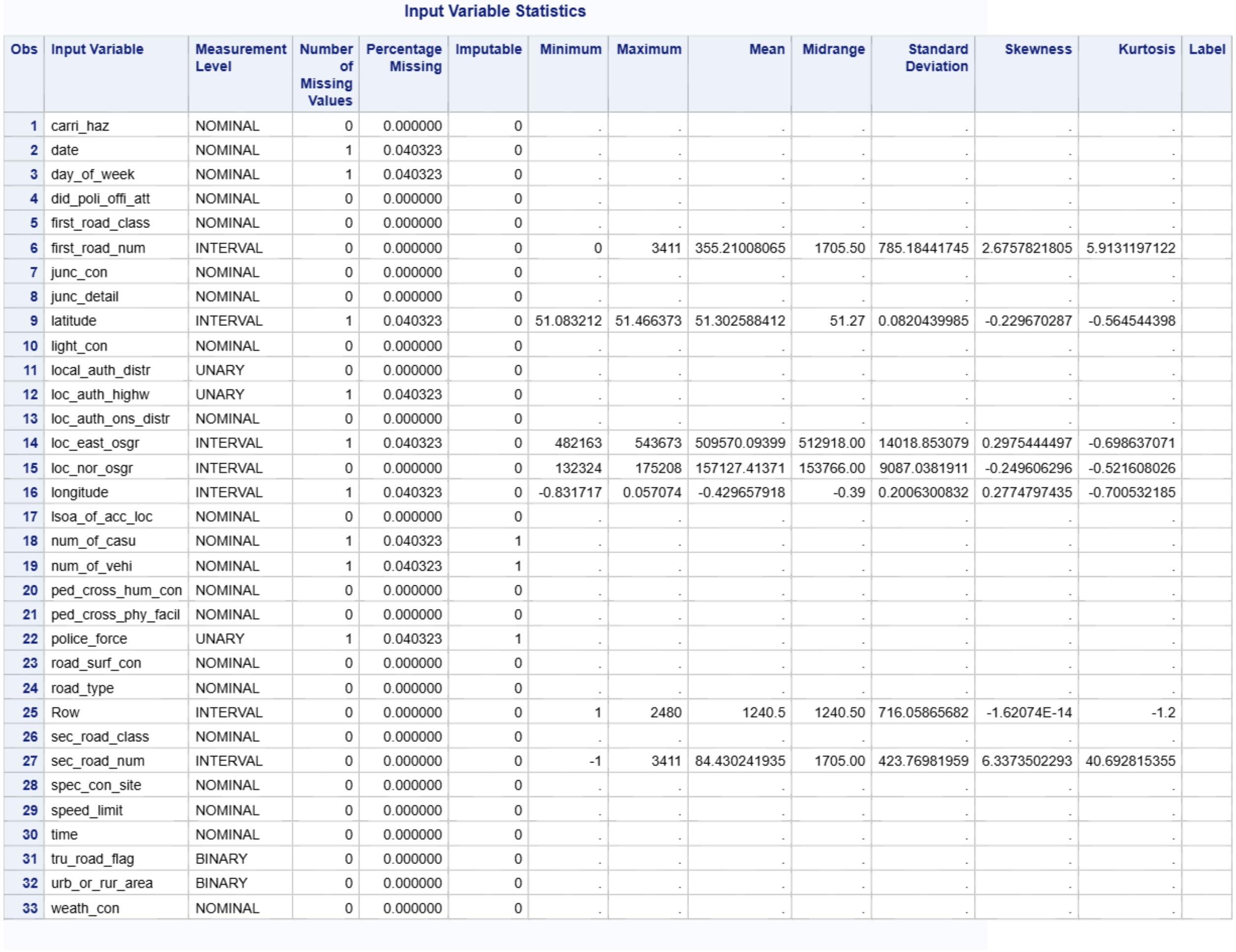
Imputed Variable	Method	Input Variable	Value	Numeric Value	Percent Missing	Variable Level	Type	Variable Label
IMP_num_of_casu	COUNT	num_of_casu		1	0.0403	NOMINAL	N	Imputed num_of_casu
IMP_num_of_vehi	COUNT	num_of_vehi		2	0.0403	NOMINAL	N	Imputed num_of_vehi
IMP_police_force	CONSTANT	police_force		0	0.0403	UNARY	N	Imputed police_force

Properties

Property Name	Property Value
bonferroni	false
codeLocation	mlearning
constantChar	
constantNum	0
dataLimit	ALLDATA
dataLimitPercent	5
dataMiningVersion	V2024.09
defClassMethod	NONE
defIntervalMethod	MEAN
fullDatasetReconstitution	false
ignoreMetadata	false
imputeNonmiss	false
indicatorRole	REJECTED
indicatorSingle	false
indicatorSubject	IMPUTED
indicatorUnique	false
intervalCrit	FTEST
leafSize	5
maxBranch	2
maxDepth	5
maxMissPercent	50
missing	USEINSEARCH
nominalCrit	CHISQUARE
partitionFraction	0.3000
prunePartition	true
pruneType	COSTCOMPLEXITY

Property Name	Property Value
randomSeed	12,345
rejectOrgVars	true
reportingOnly	false
summaryStatistics	false
templateRevision	2

Output



Input Variable Statistics

Obs	Input Variable	Measurement Level	Number of Missing Values	Percentage Missing	Imputable	Minimum	Maximum	Mean	Midrange	Standard Deviation	Skewness	Kurtosis	Label
1	carri_haz	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
2	date	NOMINAL	1	0.040323	0	.	.	.	.	.	.	.	
3	day_of_week	NOMINAL	1	0.040323	0	.	.	.	.	.	.	.	
4	did_poli_offi_att	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
5	first_road_class	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
6	first_road_num	INTERVAL	0	0.000000	0	0	3411	355.21008065	1705.50	785.18441745	2.6757821805	5.9131197122	
7	junc_con	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
8	junc_detail	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
9	latitude	INTERVAL	1	0.040323	0	51.083212	51.466373	51.302588412	51.27	0.0820439985	-0.229670287	-0.564544398	
10	light_con	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
11	local_auth_distr	UNARY	0	0.000000	0	.	.	.	.	.	.	.	
12	loc_auth_highw	UNARY	1	0.040323	0	.	.	.	.	.	.	.	
13	loc_auth_ons_distr	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
14	loc_east_osgr	INTERVAL	1	0.040323	0	482163	543673	509570.09399	512918.00	14018.853079	0.2975444497	-0.698637071	
15	loc_nor_osgr	INTERVAL	0	0.000000	0	132324	175208	157127.41371	153766.00	9087.0381911	-0.249606296	-0.521608026	
16	longitude	INTERVAL	1	0.040323	0	-0.831717	0.057074	-0.429657918	-0.39	0.2006300832	0.2774797435	-0.700532185	
17	lsoa_of_acc_loc	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
18	num_of_casu	NOMINAL	1	0.040323	1	.	.	.	.	.	.	.	
19	num_of_vehi	NOMINAL	1	0.040323	1	.	.	.	.	.	.	.	
20	ped_cross_hum_con	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
21	ped_cross_phy_facil	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
22	police_force	UNARY	1	0.040323	1	.	.	.	.	.	.	.	
23	road_surf_con	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
24	road_type	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
25	Row	INTERVAL	0	0.000000	0	1	2480	1240.5	1240.50	716.05865682	-1.62074E-14	-1.2	
26	sec_road_class	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
27	sec_road_num	INTERVAL	0	0.000000	0	-1	3411	84.430241935	1705.00	423.76981959	6.3373502293	40.692815355	
28	spec_con_site	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
29	speed_limit	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
30	time	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	
31	tru_road_flag	BINARY	0	0.000000	0	.	.	.	.	.	.	.	
32	urb_or_rur_area	BINARY	0	0.000000	0	.	.	.	.	.	.	.	
33	weath_con	NOMINAL	0	0.000000	0	.	.	.	.	.	.	.	

Road_2021_Data_Cleaning

"SAS Code" Results

by: di00222@surrey.ac.uk

Contents

Properties

Property Name	Property Value
codeLocation	mlearning
dataMiningVersion	V2024.09
exactPctlLift	true
explainFidelity	false
explainInfo	false
fullDatasetReconstitution	false
icePlots	false
maxNumShapVars	20
nBins	50
pdNumImportantInputs	5
pdObsSamples	1,000
pdPlots	false
performKernelShap	false
performLime	false
performVI	false
reportingOnly	false
sasScoreCode_Language	sas
seedId	12,345
specifyRows	RANDOM
templateRevision	2
truncateLI	5
truncateUI	95