Satyam Vatts
Avneet Kaur

PURPOSE

The Medical Field data is very crucial for analysis of diseases but often is very sparse and difficult to interpret. This report deals with a dataset consisting of medical data about difference between effects of Plasma Exchange Therapy and Vietnam's Ministry of Health's guidelines in 2015 treatments on Acute Pancreatitis. The report deals with the following analysis:

- Understanding the meaning of each variable
- Checking the Accuracy of each variable
- Selecting variables based on statistical and medical criteria
- Visualizing missing values in the data subset
- Categorizing missing data as MCAR, MAR and MNAR.

DATA DESCRIPTION

The dataset consists of 165 observations. Each observation corresponds to a patient diagnosed with hypertriglyceridemic pancreatitis and considered for the study. Among the patients, 83 were treated with PEX(Plasma Exchange Therapy) treatment while others were given treatment according to Vietnam's Ministry of Health's guidelines in 2015.

The dataset consists of 194 variables providing the complete journey of the patient throughout the hospitalization and/or until death due to the disease. Although each variable has been recorded in an effort to capture patient status, the analysis of data requires us to eliminate redundant information, non-varying parameters and missing data that can affect the prediction models.

Below is a detailed description of each of the variable in dataset. There are 29 categorical variables while rest are numerical. The table below describes the meaning of each variable, values, missing data and an explanation about its significance for analysis. Since the data consists of a lot of missing values, any variable which might be significant in study but consists of more than 50 % missing values is eliminated from further analysis. This is done to avoid biased results during regression modelling and further analysis in future.

Categorical Variables:

Column Name	Description	Data Value	Value Meaning	Freq	Prop	Decision	Reason
Gender	Patient's Gender	Nam	Male	113	0.685	Keep	Although, 68% of patients are male, the variable is required to analyse which gender is affected more.
		Nu	Female	52	0.315		
vv_reason_1	Primary reason of Hospitalization	dau bung	Stomach Ache	159	0.964	Remove	Almost everyone has same reason to admit which is stomach ache. Since there is no variablity in this data,it is not need for analysis
		Blank	Blank	6	0.036		
vv_reason_2	Breakdown of reasons of Hospitalization	dau bung thuong vi	Epigastric Abdominal Pain	87	0.527	Remove	This variable is a further explanation of main reason to admit. There is no variability in data. Only 1 category and rest is blank thus, it would not be able to act as a predictor for outcome of treatment.
		Blank	Blank	78	0.473		
vv_reason_3	Further Breakdown of reasons of Hospitalization	buon non	Nausea	2	0.012	Remove	This variable further states other symptoms patients are exhibiting. But, it contains 96% blank data i.e., most people didn't fill this reason. Thus, it can not act as a predictor of outcome of treatment
		non	Vomiting	4	0.024		
		Blank	Blank	159	0.964		
vv_others	Breakdown of reasons of Hospitalization	Dau bung man suon (P)	Abdominal Pain	1	0.006	Remove	This variable further breaks down the main reason and states what

		dau bung quanh ron	Abdominal pain around Naval	1	0.006		kind of symptoms might be causing the main reason of admission. Abdominal Pain around Naval, in general, lower flank ribs pain, shortness of breath, increase in TC etc. are most of the reasons. But, data contains 95% blank data thus, cannot act as a predictor of outcome of treatment.
		ha suon=man suon	Lower Flank Rib	1	0.006		
		kho tho	Shortness of Breadth	3	0.018		
		vtc tang triglycerid	Increase in Triglycerides	1	0.006		
		VTC tang triglycerid , gian dai be than do soi NQ	Increase in Triglycerides, pyelonephritis due to kidney stones	1	0.006		
		Blank	Blank	157	0.952		
ts_giadinh	Hereditary information	co	Yes	47	0.285	Keep	This variable specifies whether patient had a Hereditary problem or not. The missing values are 11% and can be treated in future using regression imputation since each patient is individual
		khong	No	99	0.6		
		Blank	Blank	19	0.115		
details_ts_giadinh	A breakdown of hereditary information	rl lipid	Dyslepidemia	44	0.267	Remove	The description of Hereditary disease is specified here. Although we kept the variable that specifies whether there is an hereditary issue or not, this variable providing reason does not have any variability as 60% people replied NO for Hereditary Info and 19 didn't fill, so there are 118 NA values. All other values mean Dyslipidaemia i.e., High Cholesterol.
		RLCH lipid	Dyslepidemia	1	0.006		
		RLCH lipid mau	High Rapid Metabolism	1	0.006		
		rlmm cach 2 nam	High Cholestrol from 2 years	1	0.006		
		Blank	Blank	118	0.715		
ts_benhmat	Gallbladder problem	co	Yes	1	0.006	Remove	The presence of Gallbladder problem is stated in this variable. Since, 99.4% data is a single value NO with only 1 record as yes, we cannot take it as a good predictor.
		khong	No	164	0.994		
ts_ruou	Drinking problem	co	Yes	72	0.436	Keep	Describes whether patient suffers from a drinking problem or not. Chronic alcohol consumption causes 17% to 25% of acute pancreatitis cases worldwide and is the second most common cause of AP.
		khong	No	91	0.552		
		NA	NA	1	0.006		
		Blank	Blank	1	0.006		
ts_dtd	Diabetes problem	co	Yes	32	0.194	Keep	Keep this column. It tells whether person has diabetes or not which is an important parameter in AP as diabetes is more likely to cause gallstones which is the most common cause of AP. One Invalid Value can be either treated or removed based on further analysis.
		khong	No	132	0.8		
		3	Unknown	1	0.006		
ts_vtc	Historical cholecystitis problem	co	Yes	81	0.491	Keep	Keep this column as it defines history of patient
		khong	No	81	0.491		

		3	Unknown	1	0.006		with cholecystitis(Inflammation of Gallbladder). The inflammation can be due to AP history. Two invalid values can be either imputed or removed from data.
		5	Unknown	1	0.006		
daubung	Tummy Pain	0	No	5	0.03	Remove	This field asks a query that has been already answered by patient in main reason for hospitalization. Also, it is almost similar to vv_reason_1 and contains YES as 97% of data. Thus, no variability.
		1	Yes	160	0.97		
non	Vomitting	0	No	26	0.158	Keep	Keep it as it tells patient's journey. It is one of the symptoms of AP. Those with Blanks are to be determined using variability of data.
		1	Yes	62	0.376		
		Blank	Blank	77	0.467		
ls_cn_bidaitien	Clinical symptoms of defecation	khong	No	30	0.182	Remove	This variable tells if patient displayed any clinical symptoms of constipation/obstipation. Since 72% of data is missing and 18% patient said NO as answer. Thus, data does not have variability.
		t0	At time of admission	6	0.036		
		T0	At time of admission	8	0.048		
		t30	30 hrs after admission	1	0.006		
		T30	30 hrs after admission	1	0.006		
		Blank	Blank	119	0.721		
ls_cn_ialong	Clinical symptoms of Diarrhoea	khong or 0	No	28	0.17	Remove	Clinical symptoms of Diarrhoea are seen in 7% of patients only. 17% answered NO while 75% data is blank which may also reflect no answer. Since 92% of data is similar, this variable cannot contribute to further study of disease.
		t0 or T0 or to	At time of admission	8	0.048		
		t6 or tn6 or t96	6 hrs after admission	4	0.024		
		TRV	Unknown	1	0.006		
		Blank	Blank	124	0.752		
ls_tht_bungchuong	Clinical symptoms of Abdominal distension	khong	No	4	0.024	Remove	Although it tells whether there is abdominal distension (expanded due to internal pressure) which is a common symptom in AP caused due to fluid leak into the space behind abdominal organs, it has been already covered in a sub-clinical examination cls_sa_dichob_t0 which clearly states whether patient had abdominal fluids or not. Also, almost 82% had it at time of hospitalization
		t or t0 or T0 or to	At time of admission	136	0.824		
		t0;t30;t54	Unknown	1	0.006		
		t30 or T30	30hr after admissionm	3	0.018		
		t6 or T6 or t96	6 hrs after admission	3	0.018		
		Blank	Blank	18	0.109		
ls_tt_lungsuon	Clinical symptoms of painful pressure throughout the abdomen	co or t0 or T0	Yes	3	0.018	Remove	Remove it as it is similar to Abdominal Distension. The amount of blank is 80% which might be since question has already been answered earlier. So, it is redundant data and not useful.
		khong	No	29	0.176		
		Blank	Blank	133	0.806		
cls_sa_tuy_t0		vtc or VTC	Acute Pancreatitis	83	0.503	Keep	This variable has a variety of different sub clinical

	subclinical examination - (pancreas) ultrasound at the points of admitting hospitals	vtc hoai tu	Necrotizing acute pancreatitis	7	0.042		symptoms in patients relating to AP. It tells about whether the patient already had AP, its severity and other observations through a Ultrasound of Pancreas.
		vtc phu or VTC phu	Acute Edematous Pancreatitis	30	0.182		
		VT man	Unknown	1	0.006		
		Phu or phu	Edema	8	0.048		
		han che tham kham		1	0.006		
		tham nhieu phu	Edema	1	0.006		
		khong	No	3	0.018		
		tang kt dau tuy		1	0.006		
		vuong hoi	Slightly sqaure	2	0.012		
		tang kich thuoc tham nhieu	Large inflamation	1	0.006		
		tham nhiem dau tuy	Oil Infiltration	1	0.006		
		dich quanh tuy	peripancreatic fluid	1	0.006		
		khong quan sat duoc or khong quan sat	Unobservable	2	0.012		
		Kho qs		1	0.006		
		Dich xa	Discharge	1	0.006		
		kho thay or Kho thay	Hard to see	2	0.012		
		Blank	Blank	18	0.109		
CLS_S2	miss	No Data			0	Remove	
cls_sa_dichob_t0	subclinical examination - (Abdominal fluid) ultrasound at the points of admitting hospitals	Co	Yes	110	0.667	Keep	This examination tells whether Abdominal Fuild was present at the time of hospitalization. These fluids are cause of Abdominal distension which is a common symptom of AP.
		Khong	No	36	0.218		
		Blank	Blank	19	0.115		
cls_sa_mat_t0	subclinical examination - (bladder) ultrasound at the points of admitting hospitals	0 or khong	No	2	0.006	Keep	Gallbladder ultrasound is a better test to verify gallstones which are the major cause of AP. These results help in better identifying the condition of patient and severity of AP.
		bt	Stones in Biliary Tract	78	0.473		
		polyp tu	Gallbladder Polyps History	1	0.006		
		Blank	Blank	84	0.509		
CLS_S1	miss	No Data			0	Remove	
cls_ct_tuy_lan1	subclinical examination - (pancreas) computer tomography	32mm, tham nhieu mo		1	0.006	Reomve	Covered in CTSI Score. Categories not needed for future analysis
		bo k deu, tham nhieu mo		1	0.006		
		CV≥		6	0.036		
		dich thuan nhiem quanh tuy		1	0.006		
		hoai tu		2	0.012		

		hoai tu 1 phan		1	0.006		
		Kh√¥ng		1	0.006		
		kt to,xung quanh co dich		1	0.006		
		kich thuoc k to, tham nhieu		1	0.006		
		phu		5	0.03		
		Phu		1	0.006		
		phu dich xa		1	0.006		
		phu tham nhieu mo		1	0.006		
		tang kich thuoc		1	0.006		
		phu, k hoai tu		1	0.006		
		tang kich thuoc, kem ngam thuoc		1	0.006		
		tham nhiem mo, kt bt		1	0.006		
		tham nhiem xung quanh, k hoai tu		1	0.006		
		tham nhiem, dich quanh tuy		1	0.006		
		tham nhieu dau tuy		1	0.006		
		tham nhieu mo dau tuy		1	0.006		
		tham nhieu mo quanh tuy		2	0.012		
		tham nhieu mo, tu dich sau MP		1	0.006		
		the phu		1	0.006		
		the phu VTC		1	0.006		
		to toan bo		1	0.006		
		VTC		22	0.133		
		vtc		3	0.018		
		vtc ho?i t?		1	0.006		
		vtc hoai tu		7	0.042		
		vtc phu		31	0.188		
		VTC phu		12	0.073		
		VTC phu ne		1	0.006		
		vtc the phu		5	0.03		
		VTC the phu		3	0.018		
		vtchoai tu		1	0.006		
		Blank		42	0.255		

cls_ct_dichob_lan1	subclinical examination - (Abdominal fluid) computer tomography	co or Cv= or Cv≥ or ci	Yes	69	0.418	Keep	An important aspect to determine the severity of AP. The fluid is the cause of excess pressure in abdominal area. It I released due to ill-functioning of pancreas.
		Khong or khong co or Khv¥ng	No	35	0.214		
		it	Invalid Value	2	0.012		
		2	Invalid Value	1	0.006		
		dich tu do		1	0.006		
		day 50mm		1	0.006		
		Nhieu	A lot	1	0.006		
		Blank	Blank	54	0.327		
cls_ct_balthazar_lan1	subclinical examination - balthazar score (with computer tomography)	E		42	0.255	Remove	Already covered by CTSI which is covered in numerical variables.
		e		24	0.145		
		D		23	0.139		
		d		13	0.079		
		C		11	0.067		
		c		4	0.024		
		b		2	0.012		
		A		1	0.006		
		TD VTC		1	0.006		
		Blank		44	0.267		
kq	Result - dead or alive	0	Dead	20	0.121	Keep	Required as it act as main response to treatment.
		NA	Blank	40	0.242		
		Song	Alive	105	0.636		
bcxa	Potential complication	1	Yes	79	0.479	Keep	
		NA	No	86	0.521		
pex	Patient with PEX or without PEX	1	Yes	81	0.491	Keep	Required to differentiate the two groups
		0	No	81	0.491		
		NA	Blank	3	0.018		

Numerical Variables

Column	Description	Min	Max	NA Prop	Decision	Reason

ID	Order of Observation	1	165	0	Keep	Identify Patient Individually
Age	Age of Patient	21	77	0	Keep	Required as an essential feature of patient determining contribution to disease
rv_ngaydt	Duration of staying in hospitals in days	1	18	0	Keep	Required to determine how long it took to recover or lead to death of patient
ts_ruou_nam	A breakdown of drinking problem	0	30	0.53	Keep	It provides the number of years person had drinking problem at the day of hospitalization. Since a person with a chronic alcoholism of more than 5 years is likely to manifest AP, it is an important factor for consideration. The missing data can be analysed further.
ts_ruou_nam_ml	A breakdown of drinking problem	1	1500	0.53	Remove	It is an extension of previous variable and thus is not needed.
ls_tt_alob_t0	Abdominal Pressure at time of Hospitalization	2	46	0.41	Remove	The abdominal pressure depends on the abdominal girth that can vary from patient to patient. Thus, cannot act as a good predictor.
ls_tt_bmi_t0	BMI of paitents at time of admission	15.63	31.72	0.018	Keep	The BMI reflects on patients health and is a relevant factor for determining the severity of disease. Studies have showed that obesity(BMI>25) is a major cause of AP. <18.5 : Low(Underweight) 18.5 to 24.9: Normal(Healthy) 25 to 29.9: High(Overweight) >=30 : Very High(Obese)
ls_tn_mach_t0	Heart Rate/ Pulse per minute	68	158	0.03	Keep	The admission heartrate variability acts as a significant predictor in determining AP. The normal range is 60 to 100.
ls_tn_nhiet_t0	Body temperature - Degree Celcius	36.3	39.5	0.042	Keep	Fever is a common symptom in AP and thus act as a relevant predictor for disease. Normal is 36 to 37.5 oral. Outlier Value exist in data as 366 and 3.7
ls_tn_ha_t6	Blood Pressure	90/60	140/100	0.61	Remove	Severe AP results in necrotizing pancreatitis which causes blood and pancreatic fluid to escape into the abdominal cavity, thereby decreasing blood volume. This results in a large drop in blood pressure, possibly causing shock. But the number of missing values is 60% thus, can cause biased results in prediction. Normal Range : Sys<120, Dia<80 Elevated: 120<=Sys<=139, 80<=Dia<=89 Hypertension: Sys>=140, dia>=90
ls_tn_spo2_t0	Saturation of peripheral oxygen	90	100	0.036	Keep	Presence of AP can lead to less amount of breathing due to pain. This results in low oxygen levels and thus SPO2 levels can act as a relevant predictor for the disease. Normal is 95 or Higher.
ls_tn_cvp_t0	Central Venuous Pressure	-1	30	0.73	Remove	Since, 72% of data is missing, it cannot be further utilized for analysis. Normal range is 8 to 12 mm of hg
ls_diem_apache_t0	apache 2 score at the points of admitting hospitals	0	16	0.15	Keep	The APACHE 2 score is measured to determine the severity of illness and is calculated at time of admission into ICU. It helps in determining the risk of death of patient. Score Rages from 0 to 71 depending on ICU severity.
ls_diem_ranson_t0	ranson score at the points of admitting hospitals	0	5	0.15	Keep	The Ranson Score is a scoring system that uses 11 parameters to assess the severity of AP. The 11 parameters are age, white blood cell count (WBC), blood glucose, serum aspartate transaminase (AST), serum lactate dehydrogenase (LDH), serum calcium, fall in haematocrit, arterial oxygen (PaO2), blood urea nitrogen (BUN), base deficit, and sequestration of fluids. Severity of AP. 0-2: Mortality 0to3%, 3-4: 15%, 5to6: 40%, 7 to 11: Nearly 100%. Five of the parameters should be measured after 48 hrs of admission.
ls_diem_ct_t0	CTSI score at the points of	0	10	0.32	Keep	This variable talks about Pancreatitis Severity. Score of 0-2 indicates mild, 4-6 indicates moderate and 8-10 indicates severe AP.

	admitting hospitals					
ls_diem_imrie_t0	imre score at the points of admitting hospitals	0	4	0.15	Keep	Glasgow-Imrie Criteria for Severity of AP. A score of more than 3 indicate High risk of severity of AP.
ls_diem_sofa_t0	sofa score at the points of admitting hospitals	0	7	0.15	Keep	Sequential Organ Failure Assessment score. It can be used to determine level of organ dysfunction and mortality risk in ICU patients. Since, AP leads to multiple organ failure, it is an important factor to consider 0-6 : Mortality <10% 7 to 9 : 15-20%
cls_ct_ctscore_lan1	subclinical examination - CTSI score (with computer tomography)	0	23	0.27	Remove	CTSI Score has been has already been covered in clinical examination thus not needed further.
cls_hh_bc_t0	subclinical examination - white blood cell; t0: at the points of admitting hospitals, t6: after 6h of admitting hospitals...	1.67	22.59	0.03	Keep	White Blood Cell count at time of hospitalization is an important parameter as it talks about response to an infection. Thus, a person with symptoms of AP will have increased levels of WBC and it will decrease as the patient recovers. Normal Range: 4.5 to 11 *10^9 WBC/L
cls_hh_bc_t6	subclinical examination - WBC after 6hrs	0.94	21.57	0.61	Remove	The data after 6 hrs for almost every test has around 60% of missing data. Thus, this cannot contribute to future analysis. Also, it may be because few patients may not require a WBC test after 6hrs based on condition.
cls_hh_bc_t30	subclinical examination - WBC after 30hrs	0.78	19.84	0.42	Keep	The examination of patient after 30,24 and 72 hrs are recorded after 24 hrs to check the progress of patient. It records patient journey into recovery. It is needed to verify the effect of medication being provided whether PEX or not PEX.
cls_hh_bc_t54	subclinical examination - WBC after 54hrs	2.56	20.31	0.57	Keep	
cls_hh_bc_t72	subclinical examination - WBC after 72hrs	0.5	23.68	0.55	Keep	
cls_hh_tc_t0	subclinical examination - Total Blood Count	14.6	422	0.02	Remove	The Total Blood Count is an important parameter in deciding the health of the patient. It is an accumulation of all the tests i.e., wbc, rbc, haemoglobin, haematocrit etc. Since these tests are being covered as separate variables with better accuracy, this combined score is not needed.
cls_hh_tc_t6	subclinical examination -	71	307	0.6	Remove	
cls_hh_tc_t30	subclinical examination -	1.96	296	0.44	Remove	
cls_hh_tc_t54	subclinical examination -	22	363	0.56	Remove	
cls_hh_tc_t72	subclinical examination -	52	393	0.55	Remove	
cls_hh_hct_t0	subclinical examination - Hematocrit	0.226	10.43	0.02	Keep	Haematocrit is an expression of the total percentage of blood volume that is composed of red blood cells and is also known as the packed cell volume of blood. The microcirculation disorder is the main cause of the pancreatic necrosis. There's higher vassal permeability inside the pancreatic tissue, which leads to a higher blood viscosity and its stasis in the microcirculation. Thus this test helps in detecting the AP at early stage Normal Range: Men - 41 to 50% Females - 36 to 48%
cls_hh_hct_t6	subclinical examination -	0	0.476	0.59	Remove	Since almost 60% data is missing and usually the test for HCT is done 24 hrs after to compare the severity, thus, this variable can be removed.

cls_hh_hct_t30	subclinical examination -	0.19	0.52	0.42	Keep	After 30 and 72 hrs, if the coagulation persists, the patient is severely affected. Thus, a measure of HCT is important to analyse patient's recovery over time
cls_hh_hct_t72	subclinical examination -	0.22	0.41	0.56	Keep	
cls_hh_hc_t0	red blood cell	2.7	6.88	0.12	Remove	The RBC width is an effective parameter for analysing severity of AP. RDW test is required to analyse that. Thus, this variable is not an important measure for AP cases.
cls_hh_hc_t6	subclinical examination -	2.4	468	0.63	Remove	
cls_hh_hc_t30	subclinical examination -	2.09	6.13	0.45	Remove	
cls_hh_hc_t54	subclinical examination -	2.11	5.23	0.62	Remove	
cls_hh_hc_t72	subclinical examination -	2.42	5.19	0.58	Remove	
cls_hh_pt_t0	prothrombin	36	879	0.06	Keep	It is a test to check the time it takes for blood to clot. It is seen that blood clotting is an important factor in determining severity of AP and is directly related to liver.
cls_hh_pt_t6	subclinical examination -	44	114	0.63	Remove	
cls_hh_pt_t30	subclinical examination -	50.1	6638	0.53	Keep	
cls_hh_pt_t72	subclinical examination -	0.98	134.1	0.64	Keep	
cls_hh_aptt_t0	APTT	0.76	3212	0.07	Keep	This test is similar to PT test described earlier. In this, some reagents are added in blood before checking its clotting duration. It also indicates the clotting in blood which serves a good parameter in detecting the disease
cls_hh_aptt_t6	subclinical examination -	0.41	1102	0.63	Remove	
cls_hh_aptt_t30	subclinical examination -	0.84	3.46	0.53	Keep	
cls_hh_aptt_t72	subclinical examination -	0.55	27.5	0.65	Remove	Since more than 65% data is not present, we can remove this variable.
cls_hh_fib_t0	subclinical examination - Fibrinogen	1.254	45	0.07	Keep	This is another test for blood clotting detection. It measures the amount of Fibrinogen in blood which is responsible for blood clots. Its levels indicate how the clotting system is affecting the AP condition.
cls_hh_fib_t6	subclinical examination -	1.779	8.554	0.62	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_hh_fib_t30	subclinical examination -	1.84	11.01	0.52	Keep	Enough data available for further analysis
cls_hh_fib_t72	subclinical examination -	2.56	9.292	0.64	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_sh_ure_t0	subclinical examination - ure	1.2	193	0.02	Keep	Blood Urea Nitrogen is an important parameter for severe AP. Normal Range : 6 to 24 mg/dL
cls_sh_ure_t6	subclinical examination -	1	64	0.62	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_sh_ure_t30	subclinical examination -	1.2	14.3	0.49	Keep	Although less data available, but important to observe progress of the patient. Effective after 24 hours of hospitalization.
cls_sh_ure_t72	subclinical examination -	1.3	41.9	0.55	Keep	
cls_sh_cre_t0	subclinical examination - creatinin	1.8	727	0.04	Keep	Creatinine levels increase due to AP since it relates to organ failure such as kidneys which are responsible for cleaning out creatinine from blood. Increased level are associated with AP onset.
cls_sh_cre_t6	subclinical examination -	13.5	138	0.64	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_sh_cre_t30	subclinical examination -	25	406	0.49	Keep	Although less data available, but important to observe progress of the patient.
cls_sh_cre_t72	subclinical examination -	27	328	0.55	Remove	Since its value is measure within 48hrs of admission, this variable is not important.

cls_sh_glu_t0	subclinical examination - glucose	3.2	66	0.21	Keep	The glucose levels increase during severe AP and thus prove as an important predictor for it.
cls_sh_glu_t6	subclinical examination -	5.6	64.4	0.79	Remove	Available Data is 80% or more blank, thus no insights can be gained using these variables,
cls_sh_glu_t30	subclinical examination -	5.4	28.1	0.79	Remove	Although less data available, but important to observe progress of the patient.
cls_sh_glu_t72	subclinical examination -	4.8	15.8	0.87	Remove	
cls_sh_bil_t0	subclinical examination - bilirubin total	2.1	44183	0.94	Remove	Almost No data recorded for this parameter. Thus, can be eliminated from the severity parameters.
CLS_S0	subclinical examination -	NA	NA	1	Remove	
cls_sh_bil_t6	subclinical examination -	17.1	26.7	0.99	Remove	
cls_sh_bil_t30	subclinical examination -	NA	NA	1	Remove	
cls_sh_bil_t72	subclinical examination -	5	25508	0.98	Remove	
cls_sh_gan_t0	AST, ALT (liver funtion)	45.5	45323	0.79	Remove	The data for this test is very less for making any conclusions, thus it is not considered for further analysis.
cls_sh_gan_t6	subclinical examination -	44029	44195	0.93	Remove	
cls_sh_gan_t30	subclinical examination -	16	44192	0.9	Remove	
cls_sh_ck_t0	subclinical examination -	9.78	3546	0.39	Remove	
cls_sh_chol_t0	cholesterol	3.91	99	0.18	Keep	The levels of Cholesterol HDL, LDL and Total tends to be significantly lower in patients with AP and are also associated with longer hospitalization. This data can be utilized to analyse prolongation of hospitalization of patient.
cls_sh_chol_t6	subclinical examination -	2.07	21.77	0.82	Remove	
cls_sh_chol_t30	subclinical examination -	3.1	19.8	0.79	Keep	Although less data available, but important to observe progress of the patient.
cls_sh_chol_t72	subclinical examination -	3.5	31.75	0.87	Remove	
cls_sh_tri_t0	triglycerid	11.21	131.55	0.03	Keep	It is the most important parameter for detecting severity of AP. It is the main symptom caused in patients diagnosed with the disease. Patients witness increase in Triglycerides in AP.
cls_sh_tri_t6	subclinical examination -	1.01	76.94	0.59	Remove	
cls_sh_tri_t30	subclinical examination -	0.72	84.42	0.57	Keep	Although less data available, but important to observe progress of the patient.
cls_sh_tri_t72	subclinical examination -	1.07	12.84	0.75	Remove	
cls_sh_amy_t0	subclinical examination - amylase	6.67	1519.8	0.38	Keep	Amylase is an enzyme in our blood. In Acute Pancreatitis, the level of Amylase elevates quickly after the onset of symptoms. Hence the values of the test at admission are important parameter in the diagnosis of AP.
cls_sh_amy_t6	subclinical examination -	23.6	946	0.93	Remove	It can be noticed that more than 80% of data is missing for the values taken at 6 and 30 hours. Also, the amylase level increases quickly within 12 hours of the onset of symptoms and returns to normal post that. Because of the above said reasons, we will remove these two variables.
cls_sh_amy_t30	subclinical examination -	41	860	0.84	Remove	
cls_sh_lip_t0	subclinical examination - lipase	6	1728.1	0.48	Keep	Lipase is an enzyme made by pancreas. Elevated levels of serum lipase in the test at admission support our diagnosis for AP. Hence an important parameter to keep.
cls_sh_lip_t30	subclinical examination -	30	1166	0.84	Remove	84% data is missing, therefore cannot contribute to future analysis. Also, Lipase test was done at admission to indicate AP.

cls_sh_pro_t0	subclinical examination - protein	32.1	84.3	0.24	Keep	C - Reactive Protein test. Elevated values indicate severe AP, hence Important parameter, enough data available.
cls_sh_pro_t6	subclinical examination -	47.1	70.6	0.87	Remove	More than 80% data is missing, therefore cannot contribute to future analysis.
cls_sh_pro_t54	subclinical examination -	40	71.4	0.81	Remove	
cls_sh_alb_t0	subclinical examination - albumin	13.6	56.2	0.18	Keep	Lower levels of Albumin are associated with AP, hence important parameter for our study.
cls_sh_alb_t6	subclinical examination -	2.61	43	0.84	Remove	More than 80% data is missing, therefore cannot contribute to future analysis.
cls_sh_alb_t30	subclinical examination -	1	38.9	0.83	Remove	
cls_sh_na_t0	subclinical examination - natri	4.2	154	0.05	Keep	Serum Sodium , important parameter in the diagnosis of AP
cls_sh_na_t6	subclinical examination -	122	150	0.62	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_sh_na_t30	subclinical examination -	3.7	144	0.45	Keep	Enough data available for further analysis
cls_sh_ka_t0	subclinical examination - potasium	2.6	5.3	0.07	Keep	Serum Potassium , important parameter to diagnose severity of AP
cls_sh_ka_t6	subclinical examination -	2.4	9.5	0.61	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_sh_ka_t30	subclinical examination -	2.8	137	0.44	Keep	Enough data available for further analysis
cls_sh_ka_tn6	subclinical examination -	2.33	4.5	0.79	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_sh_ca_t0	subclinical examination - calci total	0.61	35	0.76	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_ph_t0	subclinical examination - pH (in blood air)	7.1	741	0.06	Keep	arterial pH, lower values indicate higher chances of AP, important parameter for study.
cls_km_ph_t6	subclinical examination -	7.1	7.5	0.68	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_ph_t30	subclinical examination -	7.2	7.63	0.62	Keep	Although less data available, but important to observe progress of the patient.
cls_km_ph_t54	subclinical examination -	7.31	41	0.75	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_paco2_t0	subclinical examination - paCo2(in blood air)	9	97	0.07	Keep	Partial pressure of arterial carbon dioxide, important parameter to diagnose severity of AP.
cls_km_paco2_t6	subclinical examination -	14	53	0.68	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_paco2_t30	subclinical examination -	18	53	0.62	Keep	Although less data available, but important to observe progress of the patient.
cls_km_paco2_t54	subclinical examination -	20	176	0.75	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_paco2_t72	subclinical examination -	15	110.5	0.76	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_pao2_t0	subclinical examination - pa Oxy (in blood air)	32	251	0.07	Keep	Partial pressure of oxygen, important parameter to diagnose severity of AP.
cls_km_pao2_t6	subclinical examination -	45	165	0.68	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_pao2_t30	subclinical examination -	2	201	0.62	Keep	Although less data available, but important to observe progress of the patient.
cls_km_pao2_t54	subclinical examination -	16	165	0.75	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_pao2_t72	subclinical examination -	13.4	267	0.76	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.

cls_km_hco3_t0	subclinical examination - HCO3-(in blood air)	-18.6	1708	0.08	Keep	Bicarbonate , important parameter to diagnose severity of AP
cls_km_hco3_t6		5.7	224.1	0.69	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_hco3_t30		-18.9	155	0.62	Keep	Although less data available, but important to observe progress of the patient.
cls_km_hco3_t54		-12	33.5	0.75	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_hco3_t72		-11.2	39.9	0.76	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_be_t0	BE (in blood air)	-24.7	16	0.1	Keep	Base Excess in Blood Gas, important parameter to diagnose severity of AP
cls_km_be_t6		-20.2	5.6	0.7	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_be_t30		-107	10	0.64	Keep	Although less data available, but important to observe progress of the patient.
cls_km_be_t54		-13.2	390	0.75	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_be_t72		-19	333	0.78	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_pf_t0	p/f (paO2/%O2)	3.8	562	0.41	Keep	Enough data available for further analysis
cls_km_pf_t6		123	528	0.84	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_pf_t30		105	586	0.76	Keep	Although less data available, but important to observe progress of the patient.
cls_km_pf_t54		0.7	495	0.8	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_pf_t72		1.9	465	0.85	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_lac_t0	lactatr (in blood air)	0.4	9	0.11	Keep	Arterial lactate, Higher level of lactate can indicate AP, important parameter to diagnose severity of AP.
cls_km_lac_t6		0.4	5.2	0.71	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_lac_t30		0.4	4.7	0.66	Keep	Although less data available, but important to observe progress of the patient.
cls_km_lac_t54		0.4	3.2	0.75	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
cls_km_lac_t72		0.4	234	0.77	Remove	More than 50% data is missing, therefore cannot contribute to future analysis.
dt_dich_vao_t24	treatment - fluide intake	60	9650	0.1	Keep	All these vitals that are marked as 'Keep' are an important factor in comparing the PEX treatment with others. These vitals of patients help in deciding the effect of PEX Treatment and how the patient recovers over a period through this treatment. All the variables that are marked as Remove containing missing data of more than 75% of the total observations. Also, few variables denote the data of 'before the PEX medication, but the data related to 'After the PEX medication is missing, thus we need to eliminate the initial data as well. The Ranson score although is very important, but the data is available for only 2 patients after PEX. Thus, cannot be utilized for analysis.
dt_dich_vao_t48	treatment - fluide intake	1000	9500	0.14	Keep	
dt_dich_vao_t72	treatment - fluide intake	1200	8500	0.24	Keep	
dt_dich_ra_t24	treatment - fluide output	950	6900	0.09	Keep	
dt_dich_ra_t48	treatment - fluide output	620	10760	0.13	Keep	
dt_dich_ra_t72	treatment - fluide output	270	8020	0.24	Keep	
dt_dich_bilan_t24	treatment - balance fluid in and out	-2100	23200	0.11	Keep	
dt_dich_bilan_t48	treatment - balance fluid in and out	-4550	6830	0.16	Keep	
dt_dich_bilan_t72	treatment - balance fluid in and out	-3780	2650	0.26	Keep	

dt_nhin_ngay	treatment - day without food intake	0	12	0.06	Keep
dt_pex_ngaybenh	treatment - PEX treatment of which day of the diagnosis	1	7	0.51	Keep
dt_pex_lan	treatment - number of PEX treatment	1	3	0.5	Keep
dt_pex_sauvv	treatment - PEX treatment after of how many hours of the diagnosis	4	41	0.61	Keep
DT_PE0	treatment	NA	NA	1	Remove
dt_pex_tri_t_lan1	treatment - triglycerid before first time of PEX	2.41	131.55	0.5	Keep
dt_pex_tri_s_lan1	treatment - triglycerid after first time of PEX	1.01	76.94	0.52	Keep
dt_pex_chol_t_lan1	treatment - cholesterol before first time of PEX	1.14	135.13	0.56	Keep
dt_pex_chol_s_lan1	treatment - cholesterol after first time PEX	2.07	21.77	0.73	Keep
dt_pex_ldl_t_lan1	treatment - LDL before first time of PEX	0.1	11.24	0.81	Remove
dt_pex_ldl_s_lan1	treatment - LDL after first time of PEX	0.37	5.46	0.9	Remove
dt_pex_hdl_t_lan1	treatment - HDL - before first time PEX	0	9.05	0.73	Remove
dt_pex_apache_t_lan1	treatment - APAche 2 score before first time PEX	0	16	0.5	Keep
dt_pex_apache_s_lan1	treatment - APAche 2 score after first time PEX	0	9	0.52	Keep
dt_pex_ranson_t_lan1	treatment - ranson score before first time PEX	0	5	0.5	Remove
dt_pex_ranson_s_lan1	treatment - ranson score after first time PEX	2	3	0.99	Remove
dt_pex_imrie_t_lan1	treatment - Imre score before first time of PEX	0	4	0.5	Keep
dt_pex_imrie_s_lan1	treatment - Imre score after first time of PEX	0	3	0.52	Keep
dt_pex_balthazar_t_lan1	treatment - balthazar score (with computer tomography) before first time PEX	0	10	0.61	Remove

dt_pex_balthazar_s_lan1	treatment - balthazar score (with computer tomography) after first time PEX	3	6	0.96	Remove
dt_pex_sofa_t_lan1	treatment - sofa score before first time of PEX	0	7	0.5	Keep
dt_pex_sofa_s_lan1	treatment - sofa score after first time of PEX	0	8	0.54	Keep
dt_pex_alob_t_lan1	treatment - Abdominal pressure before first time of PEX	6	46	0.66	Keep
dt_pex_alob_s_lan1	treatment - Abdominal pressure after first time of PEX	5	33	0.73	Keep
kq	Result - dead or alive	0	1	0.24	Keep
bcxa	Potential complication	1	1	0.52	Keep
pex	Patient with PEX or without PEX	0	1	0.02	Keep

Accuracy Check

The variables contain a lot of NULL values as either 'NA' in numerical data or a Blank in categorical data. The proportions of these are given in the tables above. Moreover, on performing accuracy check on available values all the variables, the following observations are made:

- details_ts_giadinh : A breakdown of hereditary information contains values rl lipid, RLCH lipid, RLCH lipid mau, rlmm cach 2 nam that all mean Dyslepidemia which is High Cholesterol. Thus, these values can be combined to a single value. Since, all the responses are same, this variable has been removed from the compressed dataset.
- There are some numerical values like 0,3,5 present in some of the categorical variables. Since there is no mapping available, it is not possible to decode the meaning of these values. Thus, these are to be considered as Null Values.
- ts_vtc_lancuoi : Last detection of cholecystitis problem is a date value. But it should be removed because we do not have reference value to calculate the data where relative dates are provided.
- There are also many variables of categorical type that contain values with different cases like t0, T0, vtc, VTC etc. These values are to be clubbed together to eliminate inconsistencies in the dataset. All such values are mentioned above in the categorical variables table.
- The numerical variables relate to various tests and thus have specified ranges that mark their accuracy. Below is a list of variables along with associated inaccuracies found in the dataset.
    - ls_tn_cvp_t0 : Central Venuous Pressure -1, 0 and 99 are invalid values. But since we are not keeping it in our subset, we do not need to address them
    - cls_ct_ctscore_lan1: subclinical examination - CTSI score (with computer tomography) has two invalid values e and 23. These are not valid since the range of score is from 0 to 10
    - cls_hh_hct_t0 : subclinical examination of Haematocrit has two invalid values 3 and 10.43. These values are not correct since it can range between 0 and 1 as a proportion.
    - cls_hh_hc_t6: Subclinical Examination of red blood cells has an outlier value of 468. It must be a two-digit number as normal range varies between 4.35 to 5.65.
    - cls_hh_pt_t30 : Subclinical exam of Prothrombin has an invalid value of 6638 which cannot be possible.
    - cls_hh_aptt_t0 and t6: Subclinical exam of APTT have invalid values of 3212 & 1102.
    - cls_sh_ure_t0: subclinical examination – urea has an invalid value of 193 which very far from normal range.
    - cls_sh_bil_t0: subclinical examination - bilirubin total has a lot of invalid values. It is because these values are given as a fraction value and thus some values are missing the fraction symbol resulting in erroneous values. Similar is with variables cls_sh_gan_t0,t6,t30,t54 etc.
    - cls_sh_chol_t0 : This variable that depicts cholesterol levels also has an invalid value of 99.
    - cls_km_hco3_t0: HCO3 in blood air is being denoted by this variable recorded at various intervals. The values associated have few negative data and some values are very large which are out of the measurable range. These values must be addressed before proceeding with any analysis based on these variables.

Data Subset

Below is the final data subset taken after eliminating non-required columns. It consists of these 98 columns with all the 165 observations.

Column	**Description**
ID	Order of observations

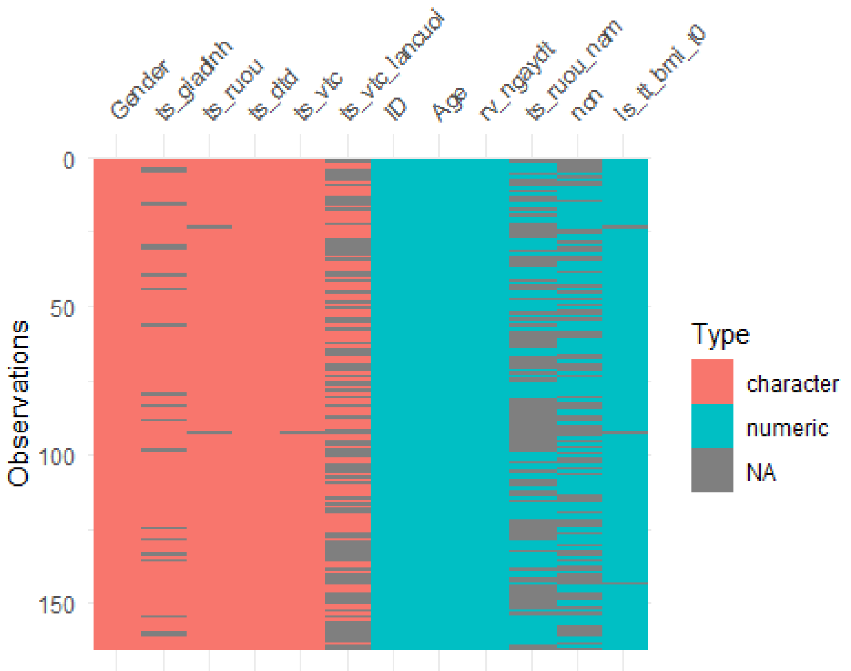
Age	Age of the Patient
Gender	Gender of Patient
rv_ngaydt	Duration of staying in hospitals
ts_giadinh	Hereditary information
ts_ruou	Drinking problem
ts_ruou_nam	A breakdown of drinking problem
ts_dtd	Diabetes problem
ts_vtc	Historical cholecystitis problem
ts_vtc_lancuoi	Last detection of cholecystitis problem
non	Vomiting
ls_tt_bmi_t0	Clinical symptoms of BMI
ls_tn_mach_t0	Clinical symptoms of Heat Rate or Pulse per Rate
ls_tn_nhiet_t0	Body temperature
ls_tn_spo2_t0	Saturation of peripheral oxygen
ls_diem_apache_t0	APACHE 2 score at the points of admitting hospitals
ls_diem_ranson_t0	RANSON score at the points of admitting hospitals
ls_diem_ct_t0	CTSI score at the points of admitting hospitals
ls_diem_imrie_t0	IMRE score at the points of admitting hospitals
ls_diem_sofa_t0	SOFA score at the points of admitting hospitals
cls_sa_tuy_t0	subclinical examination - (pancreas) ultrasound at the points of admitting hospitals
cls_sa_dichob_t0	subclinical examination - (Abdominal fluid) ultrasound at the points of admitting hospitals
cls_sa_mat_t0	subclinical examination - (bladder) ultrasound at the points of admitting hospitals
cls_ct_dichob_lan1	subclinical examination - (Abdominal fluid) computer tomography
cls_hh_bc_t0	subclinical examination - white blood cell; t0: at the points of admitting hospitals, t6: after 6h of admitting hospitals...
cls_hh_bc_t30	subclinical examination – WBC after 30hrs
cls_hh_bc_t54	subclinical examination – WBC after 54 hrs
cls_hh_bc_t72	subclinical examination – WBC after 72hrs
cls_hh_hct_t0	subclinical examination – Hematocrit
cls_hh_hct_t30	subclinical examination – HT after 30 hrs
cls_hh_hct_t72	subclinical examination – HT after 72 hrs
cls_hh_pt_t0	Prothrombin Test at time of hospitalization
cls_hh_pt_t30	subclinical examination – PT after 30hrs
cls_hh_pt_t72	subclinical examination – PT after 72hrs
cls_hh_aptt_t0	APTT Test at time of hospitalization
cls_hh_aptt_t30	subclinical examination – APTT after 30 hrs
cls_hh_fib_t0	subclinical examination – Fibrinogen Test at time of hospitalization
cls_hh_fib_t30	subclinical examination -FT after 30hrs
cls_sh_ure_t0	subclinical examination – UREA Test at time of hospitalization
cls_sh_ure_t30	subclinical examination – UREA after 30hrs
cls_sh_ure_t72	subclinical examination – UREA after 72 hrs
cls_sh_cre_t0	subclinical examination – Creatinine Test at time of hospitalization
cls_sh_cre_t30	subclinical examination -CRE after 30hrs
cls_sh_glu_t0	subclinical examination – Glucose Test for Blood Sugar at time of hospitalization
cls_sh_chol_t0	Cholesterol Test at time of hospitalization
cls_sh_chol_t30	subclinical examination – Chol after 30hrs
cls_sh_tri_t0	Triglyceride Test at time of hospitalization
cls_sh_tri_t30	subclinical examination – Tri after 30hrs
cls_sh_amy_t0	subclinical examination – Amylase Test at time of hospitalization
cls_sh_lip_t0	subclinical examination – Lipase Test at time of hospitalization
cls_sh_pro_t0	subclinical examination –Protein Test at time of hospitalization
cls_sh_alb_t0	subclinical examination – Albumin Test at time of hospitalization
cls_sh_na_t0	subclinical examination – NATRI Test at time of hospitalization
cls_sh_na_t30	subclinical examination – NATRI Test after 30hrs
cls_sh_ka_t0	subclinical examination – Potassium Test at time of hospitalization
cls_sh_ka_t30	subclinical examination – Potassium Test after 30hrs
cls_km_ph_t0	subclinical examination - pH (in blood air) Test at time of hospitalization
cls_km_ph_t30	subclinical examination –ph in Blood Air after 30 hrs
cls_km_paco2_t0	subclinical examination - paCo2(in blood air) Test at time of hospitalization
cls_km_paco2_t30	subclinical examination – paCo2 after 30hrs
cls_km_pao2_t0	subclinical examination - pa Oxy (in blood air) Test at time of hospitalization

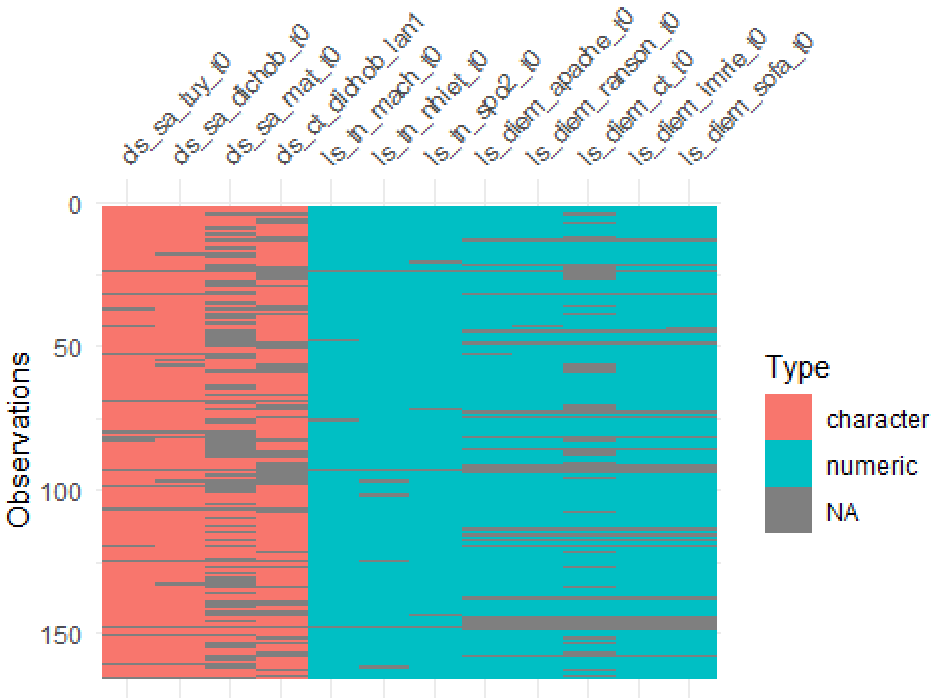
cls_km_pao2_t30	subclinical examination -pa Oxy after 30hrs
cls_km_hco3_t0	subclinical examination - HCO3-(in blood air) Test at time of hospitalization
cls_km_hco3_t30	HCO3 after 30hrs
cls_km_be_t0	BE (in blood air) Test at time of hospitalization
cls_km_be_t30	BE (in blood air) after 30hrs
cls_km_pf_t0	p/f (paO2/%O2) Test at time of hospitalization
cls_km_pf_t30	p/f (paO2/%O2) after 30hrs
cls_km_lac_t0	Lactatr (in blood air) at time of hospitalization
cls_km_lac_t30	Lactatr (in blood air) after 30hrs
dt_dich_vao_t24	treatment - fluide intake after 24 hrs
dt_dich_vao_t48	treatment - fluide intake after 48 hrs
dt_dich_vao_t72	treatment - fluide intake after 72 hrs
dt_dich_ra_t24	treatment - fluide output after 24 hrs
dt_dich_ra_t48	treatment - fluide output after 48 hrs
dt_dich_ra_t72	treatment - fluide output after 72 hrs
dt_dich_bilan_t24	treatment - balance fluid in and out after 24 hrs
dt_dich_bilan_t48	treatment - balance fluid in and out after 48 hrs
dt_dich_bilan_t72	treatment - balance fluid in and out after 72 hrs
dt_nhin_ngay	treatment - day without food intake
dt_pex_ngaybenh	treatment - PEX treatment of which day of the diagnosis
dt_pex_lan	treatment - number of PEX treatment
dt_pex_sauvv	treatment - PEX treatment after of how many hours of the diagnosis
dt_pex_tri_t_lan1	treatment - triglycerid before first time of PEX
dt_pex_tri_s_lan1	treatment - triglycerid after first time of PEX
dt_pex_chol_t_lan1	treatment - cholesterol before first time of PEX
dt_pex_chol_s_lan1	treatment - cholesterol after first time PEX
dt_pex_apache_t_lan1	treatment - APAche 2 score before first time PEX
dt_pex_apache_s_lan1	treatment - APAche 2 score after first time PEX
dt_pex_imrie_t_lan1	treatment - Imre score before first time of PEX
dt_pex_imrie_s_lan1	treatment - Imre score after first time of PEX
dt_pex_sofa_t_lan1	treatment - sofa score before first time of PEX
dt_pex_sofa_s_lan1	treatment - sofa score after first time of PEX
dt_pex_alob_t_lan1	treatment - Abdominal pressure before first time of PEX
dt_pex_alob_s_lan1	treatment - Abdominal pressure after first time of PEX
kq	Result - dead or alive
bcxa	Potential complication
pex	Patient with PEX or without PEX

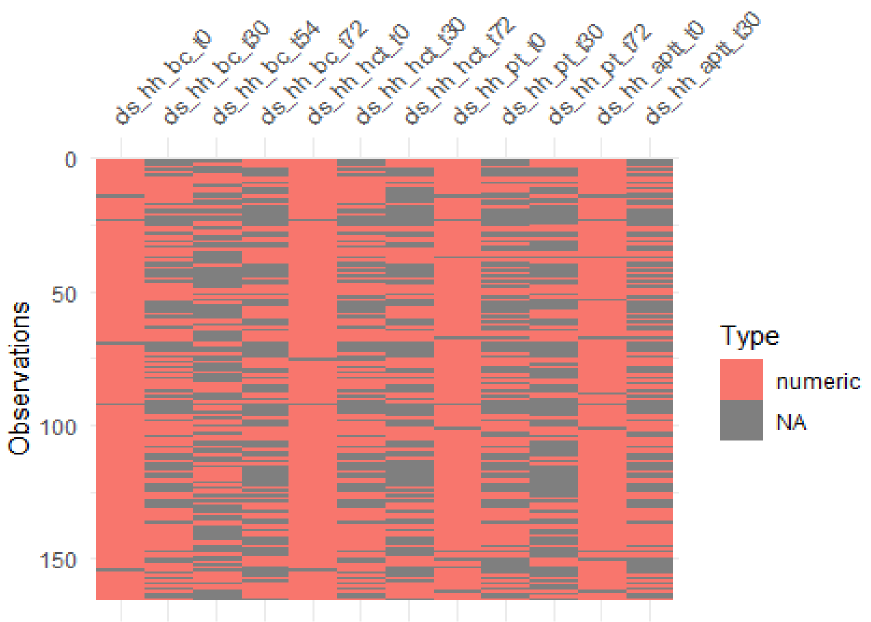
Missing Data Visualization

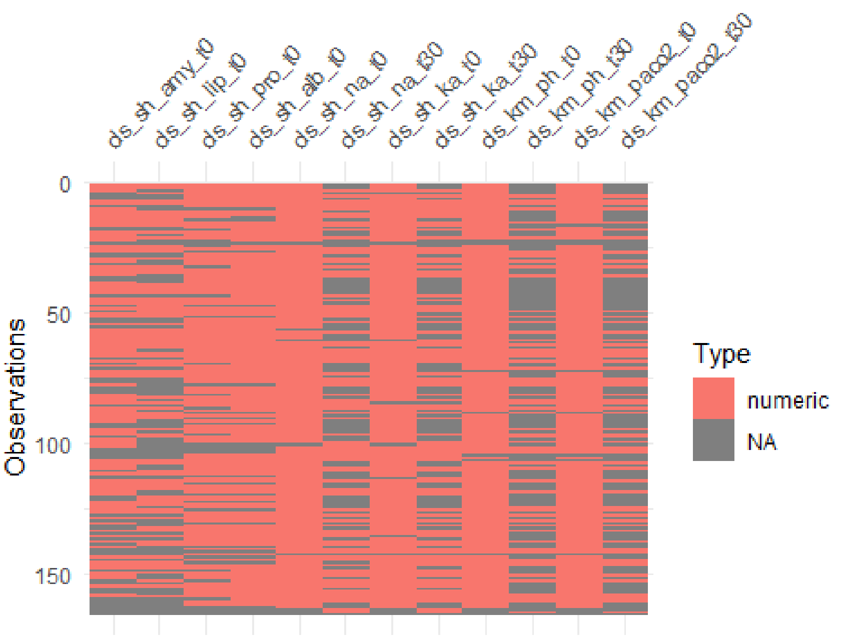
The missing data is hard to visualize using normal bar charts or tabular data. The patterns and similarities between missing data of different variables in a dataset can be easily captured by below given charts. Each chart has the dataset fields on x-axis and the order of observation on y-aixs. Presence of value is denoted by empty space while null values are being denoted by grey bars. The width of the bar is based on the frequency of missing data. Since here we are visualizing row-wise data for each column simultaneously, it is extremely easy to identify patterns in the missing data.

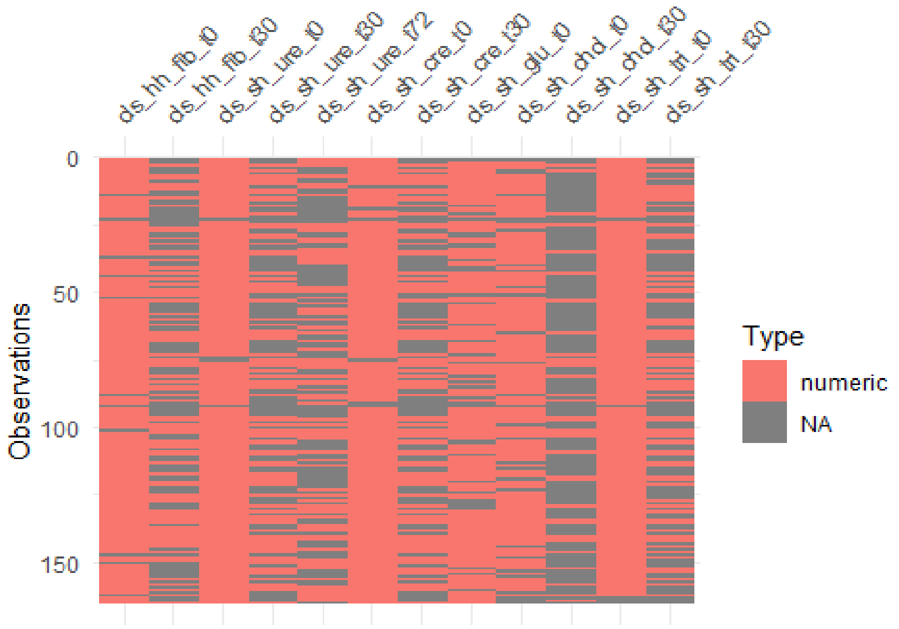
Firstly, the plots are developed based on type of data i.e., Character and Numeric. This enables us to identify whether one type of data is related to other type in terms of missing values.

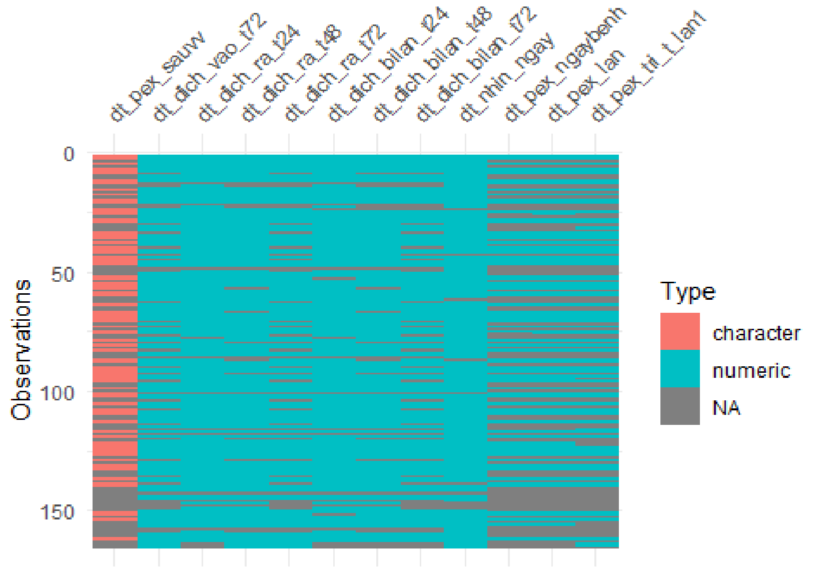

Gender
ts_giadinh
ts_ruou
ts_dtd
ts_vtc
ts_vtc_lancuoi
ID
Age
rv_ngaydt
ts_ruou_nam
non
ls_tt_bmi_t0
Observations
0
50
100
150
Type
character
numeric
NA

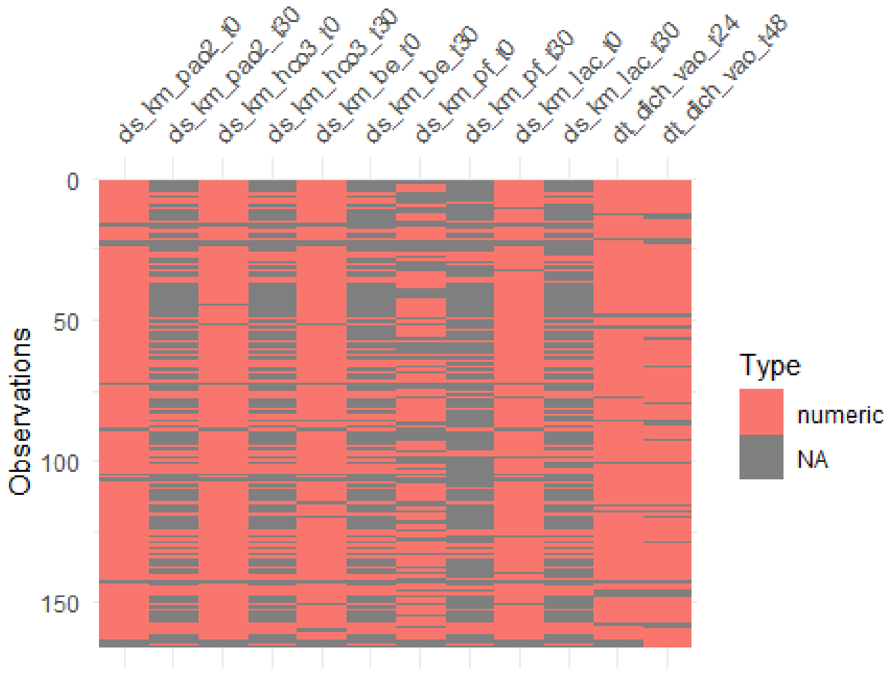

ds_sa_tuy_t0
ds_sa_dichob_t0
ds_sa_mat_t0
ds_ct_dichob_lan1
ls_tn_mach_t0
ls_tn_nhiet_t0
ls_tn_spo2_t0
ls_diem_apache_t0
ls_diem_ranson_t0
ls_diem_ct_t0
ls_diem_imrie_t0
ls_diem_sofa_t0
Observations
0
50
100
150
Type
character
numeric
NA

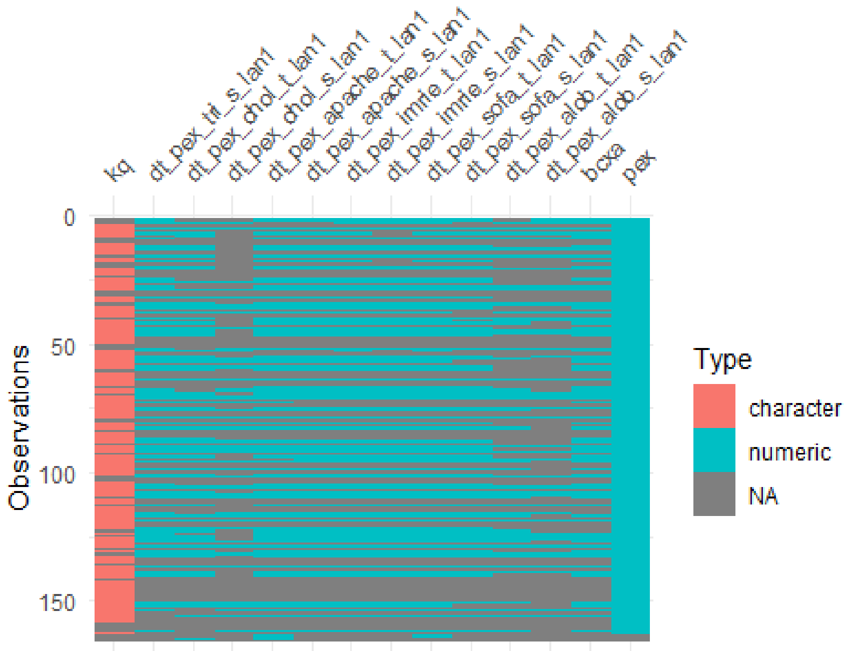

ds_hh_bc_t0
ds_hh_bc_t30
ds_hh_bc_t54
ds_hh_bc_t72
ds_hh_hct_t0
ds_hh_hct_t30
ds_hh_hct_t72
ds_hh_pt_t0
ds_hh_pt_t30
ds_hh_pt_t72
ds_hh_aptt_t0
ds_hh_aptt_t30
Observations
0
50
100
150
Type
numeric
NA


ds_sh_amy_t0
ds_sh_lip_t0
ds_sh_pro_t0
ds_sh_alb_t0
ds_sh_na_t0
ds_sh_na_t30
ds_sh_ka_t0
ds_sh_ka_t30
ds_km_ph_t0
ds_km_ph_t30
ds_km_paco2_t0
ds_km_paco2_t30
Observations
0
50
100
150
Type
numeric
NA

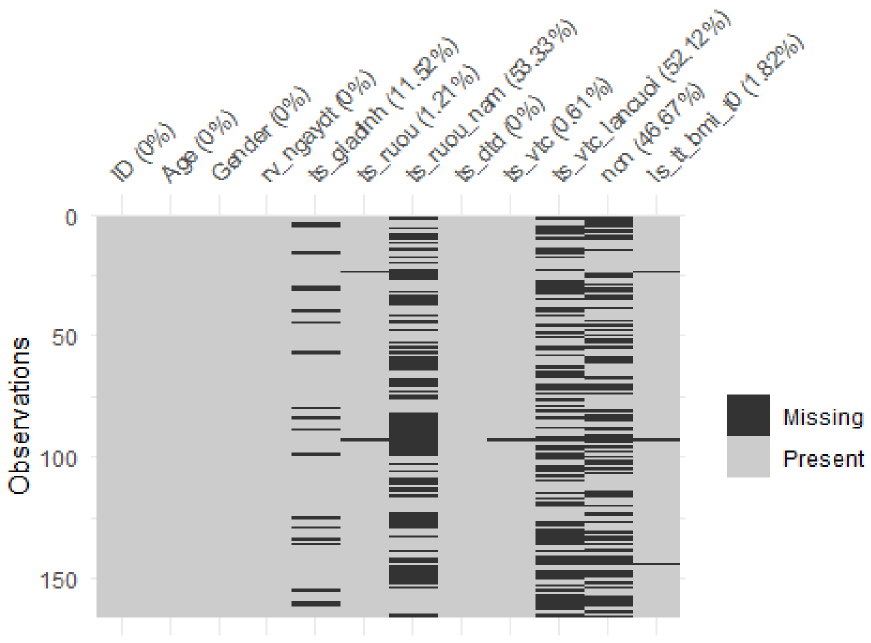

ds_hh_fib_t0
ds_hh_fib_t30
ds_sh_ure_t0
ds_sh_ure_t30
ds_sh_ure_t72
ds_sh_cre_t0
ds_sh_cre_t30
ds_sh_glu_t0
ds_sh_chol_t0
ds_sh_chol_t30
ds_sh_tri_t0
ds_sh_tri_t30
Observations
0
50
100
150
Type
numeric
NA

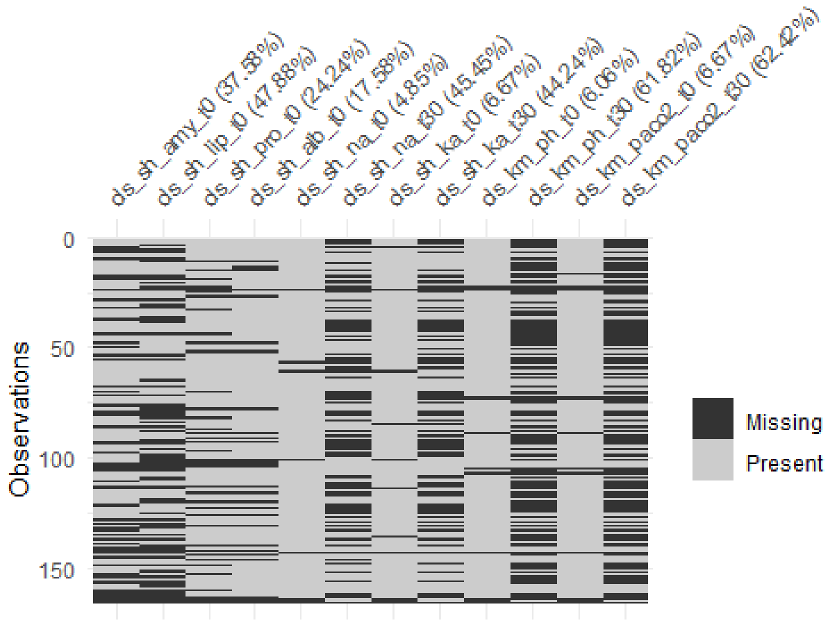

dt_pex_sauvv
dt_dich_vao_t72
dt_dich_ra_t24
dt_dich_ra_t48
dt_dich_ra_t72
dt_dich_bilan_t24
dt_dich_bilan_t48
dt_dich_bilan_t72
dt_nhin_ngay
dt_pex_ngaybenh
dt_pex_lan
dt_pex_tri_t_lan1
Observations
0
50
100
150
Type
character
numeric
NA

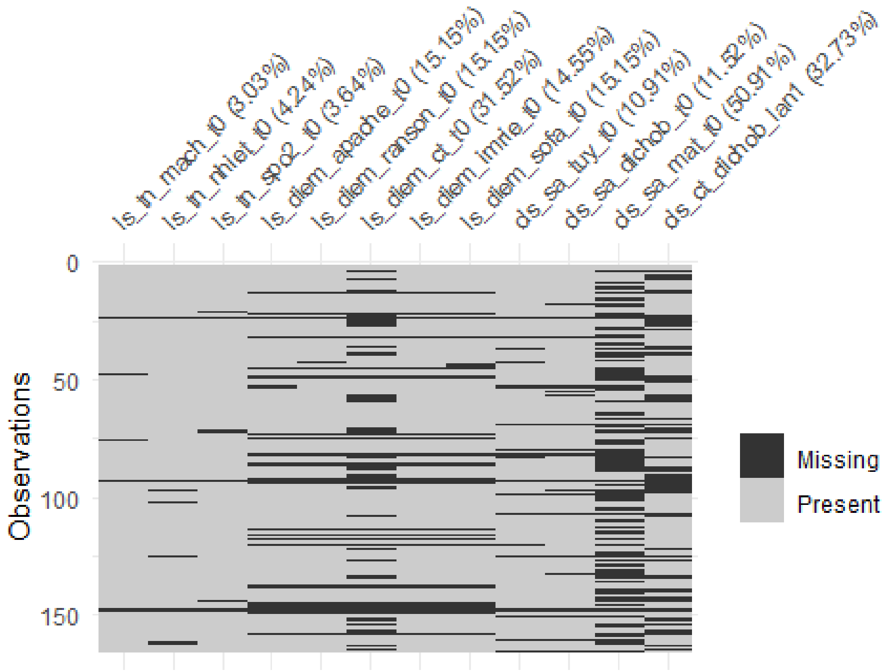


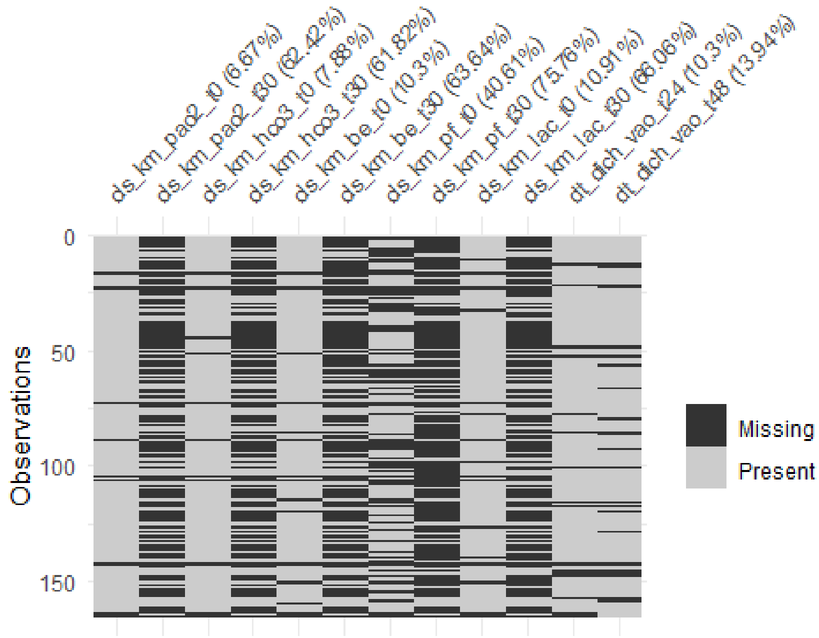


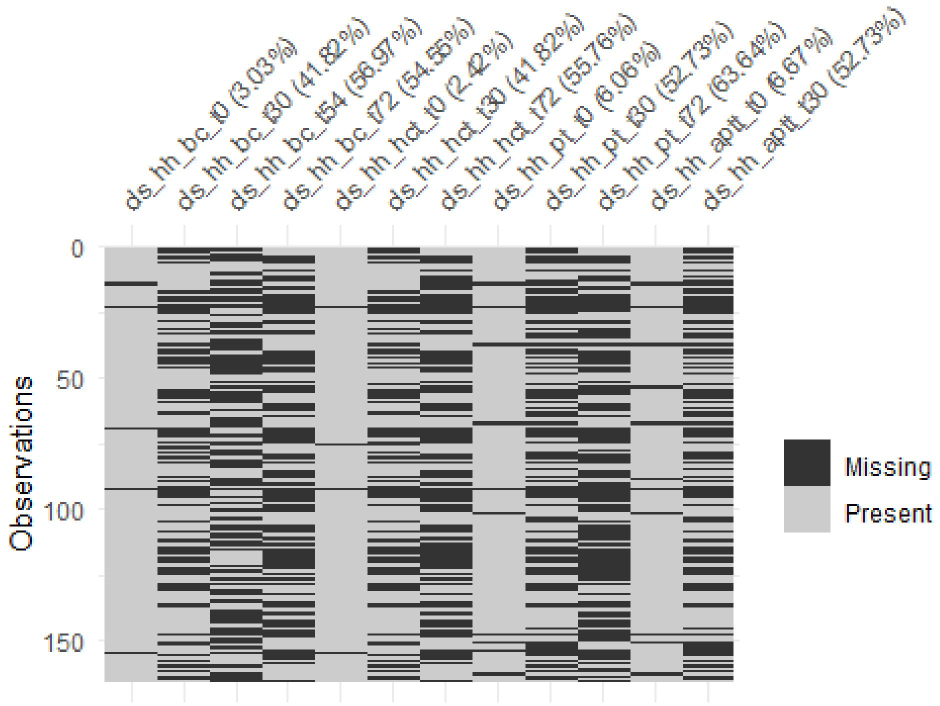
The below given plots missing and present data alongwith the oercentage of missing value. This allows s to compare different varaibles and identify relationships.

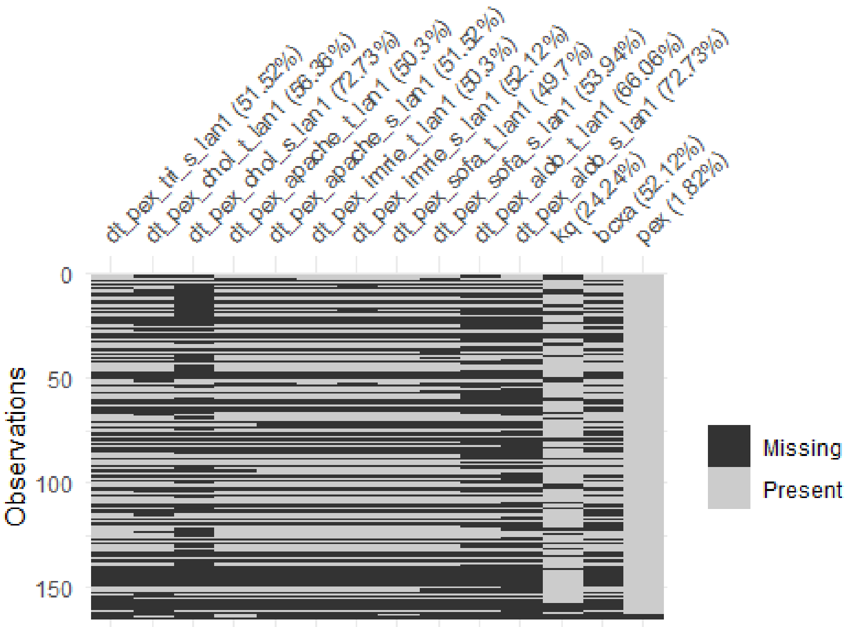


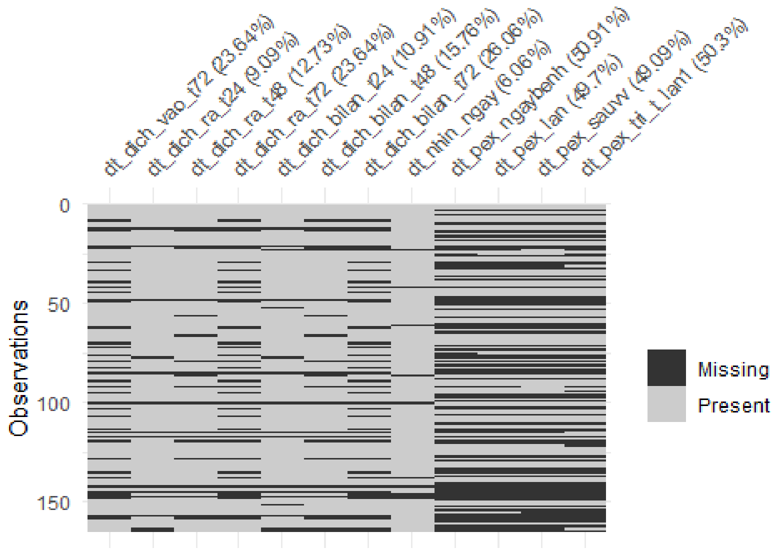


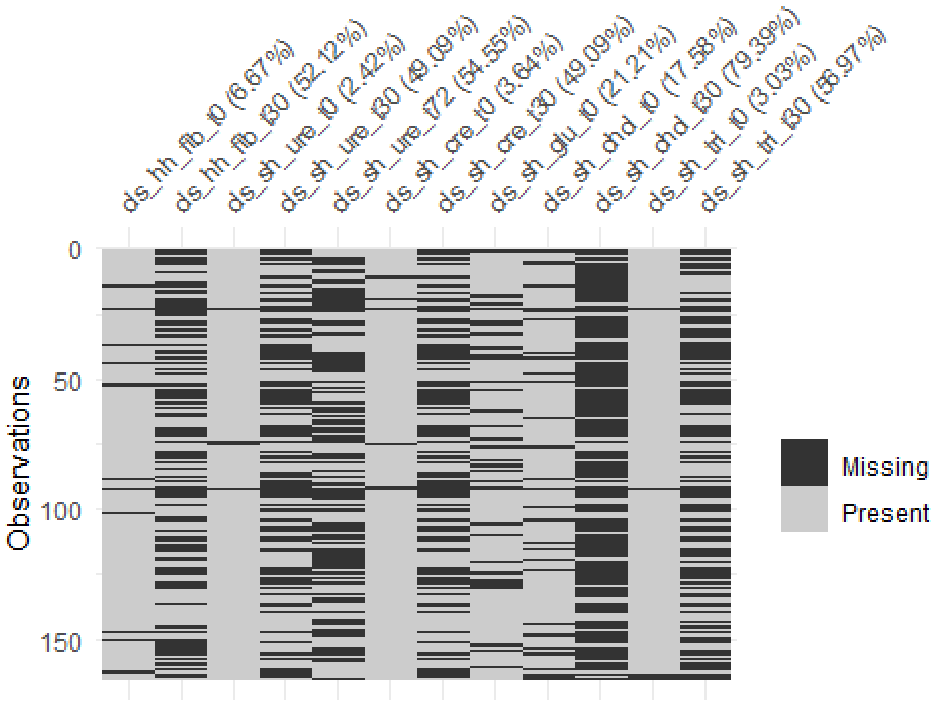




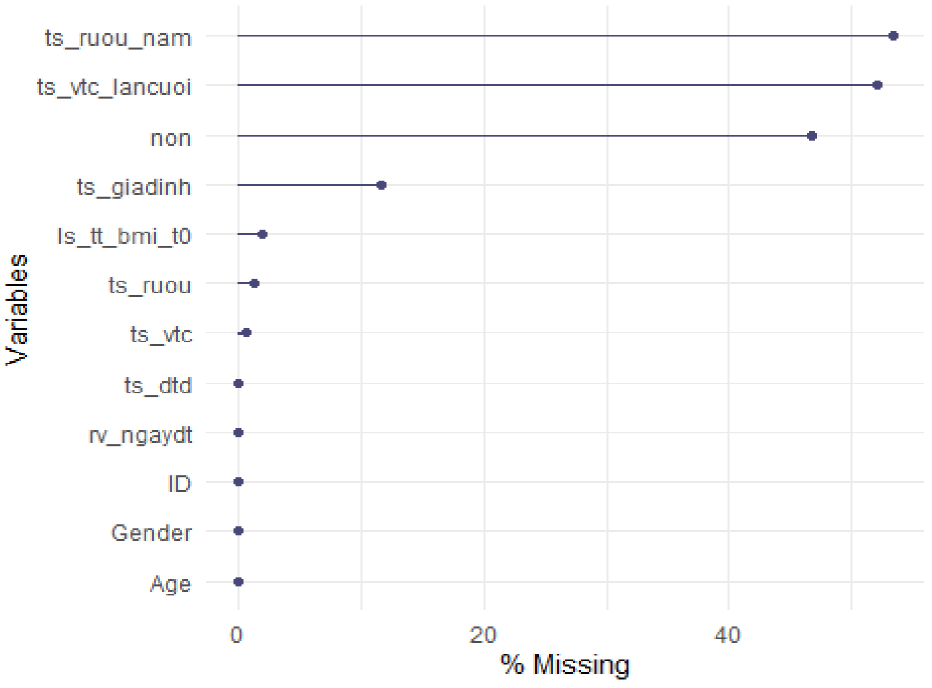


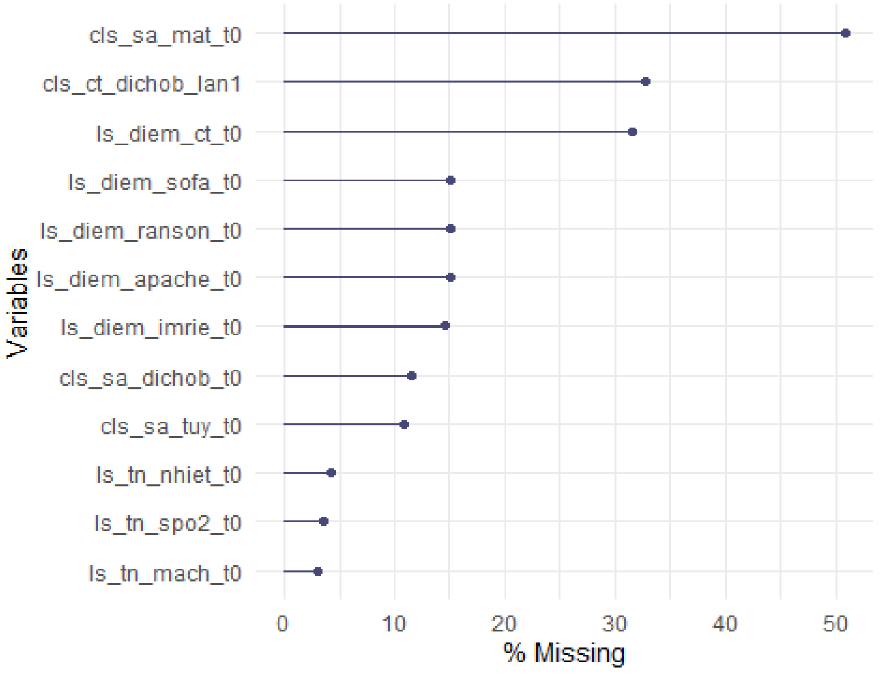


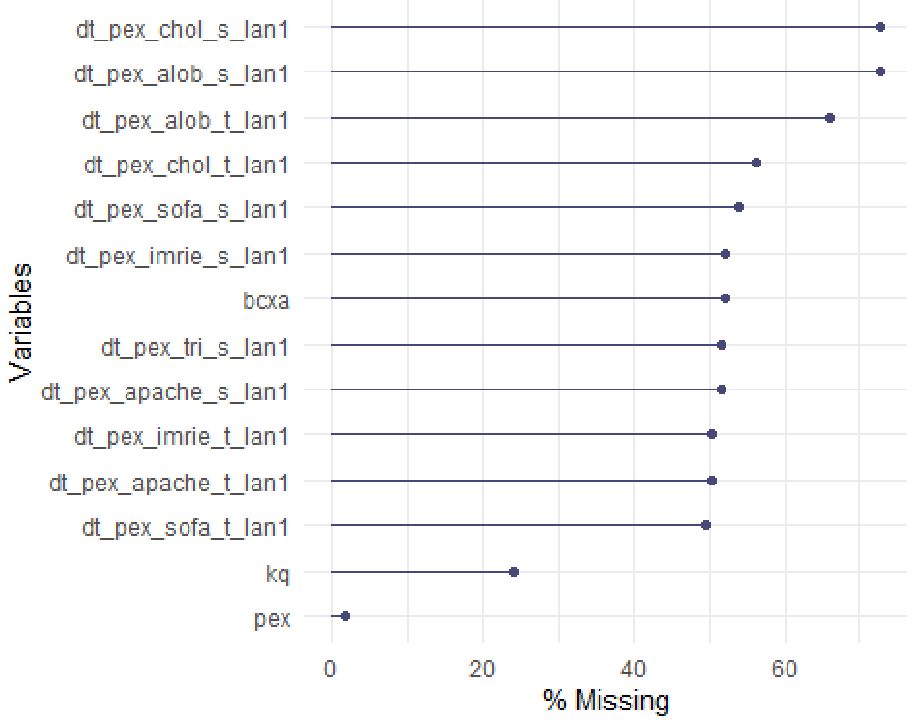


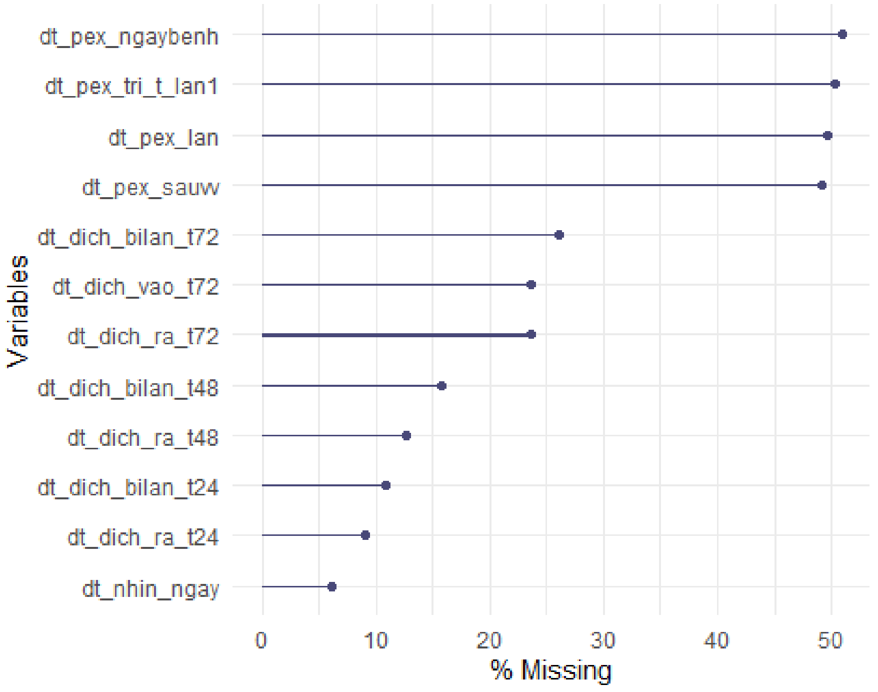


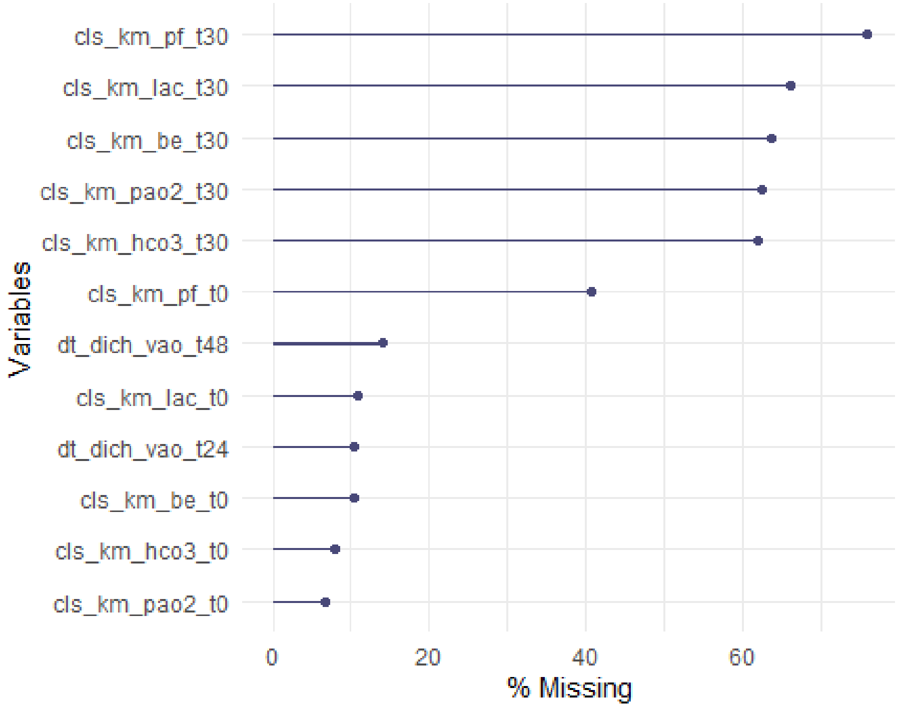
The plots listed below compares the amount of missing data for each variable on a fixed scale in order to determine how much data is missing and for which variable. It enables us to analyse the irrelevant data that might create bias. The x-axis denotes percentage missing data while y-axis has all the different variables of the data. The lines across the plot signify the percentage of missing data.

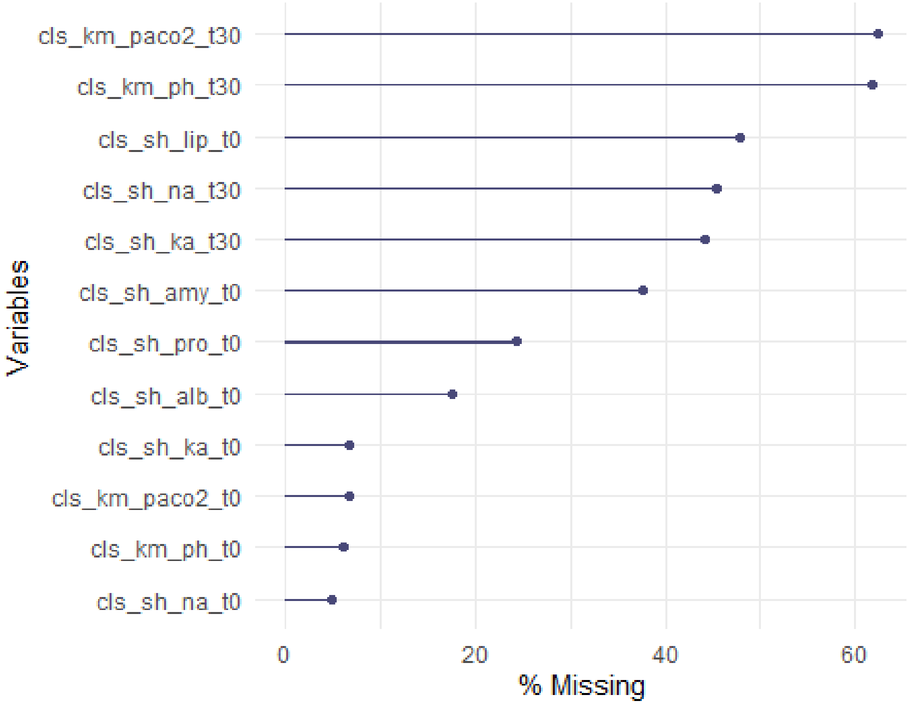


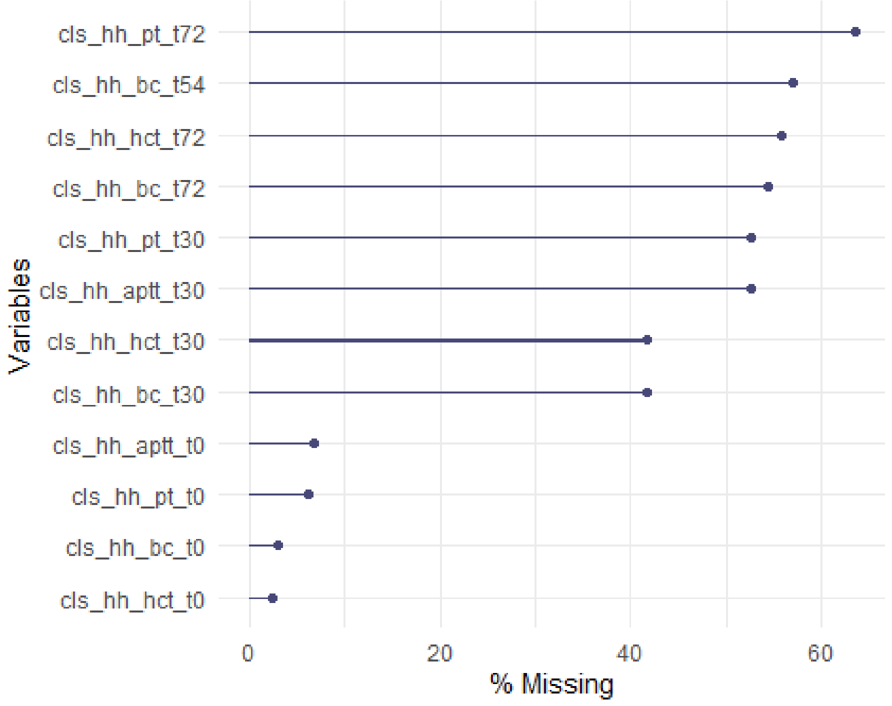


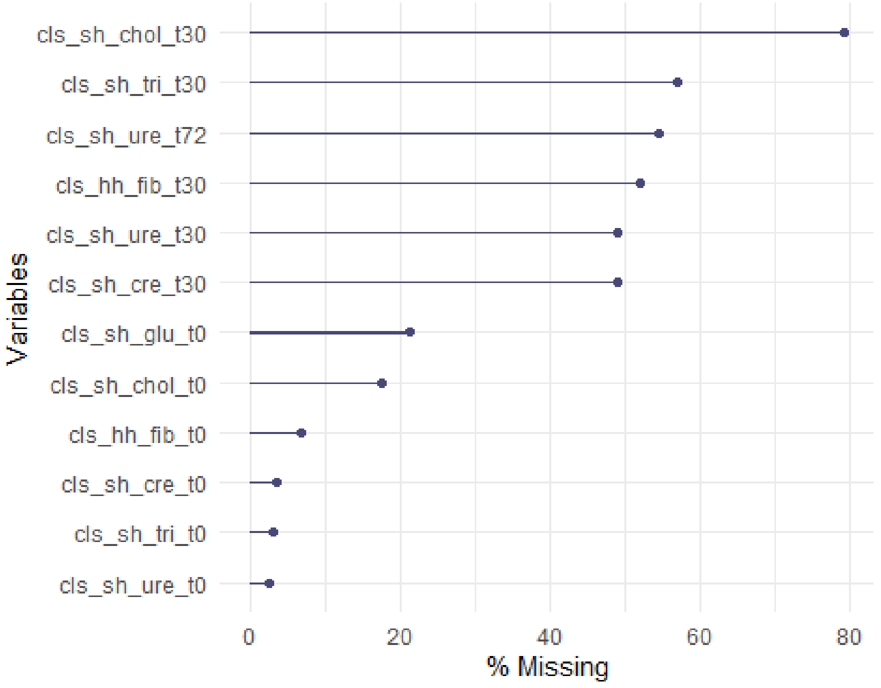






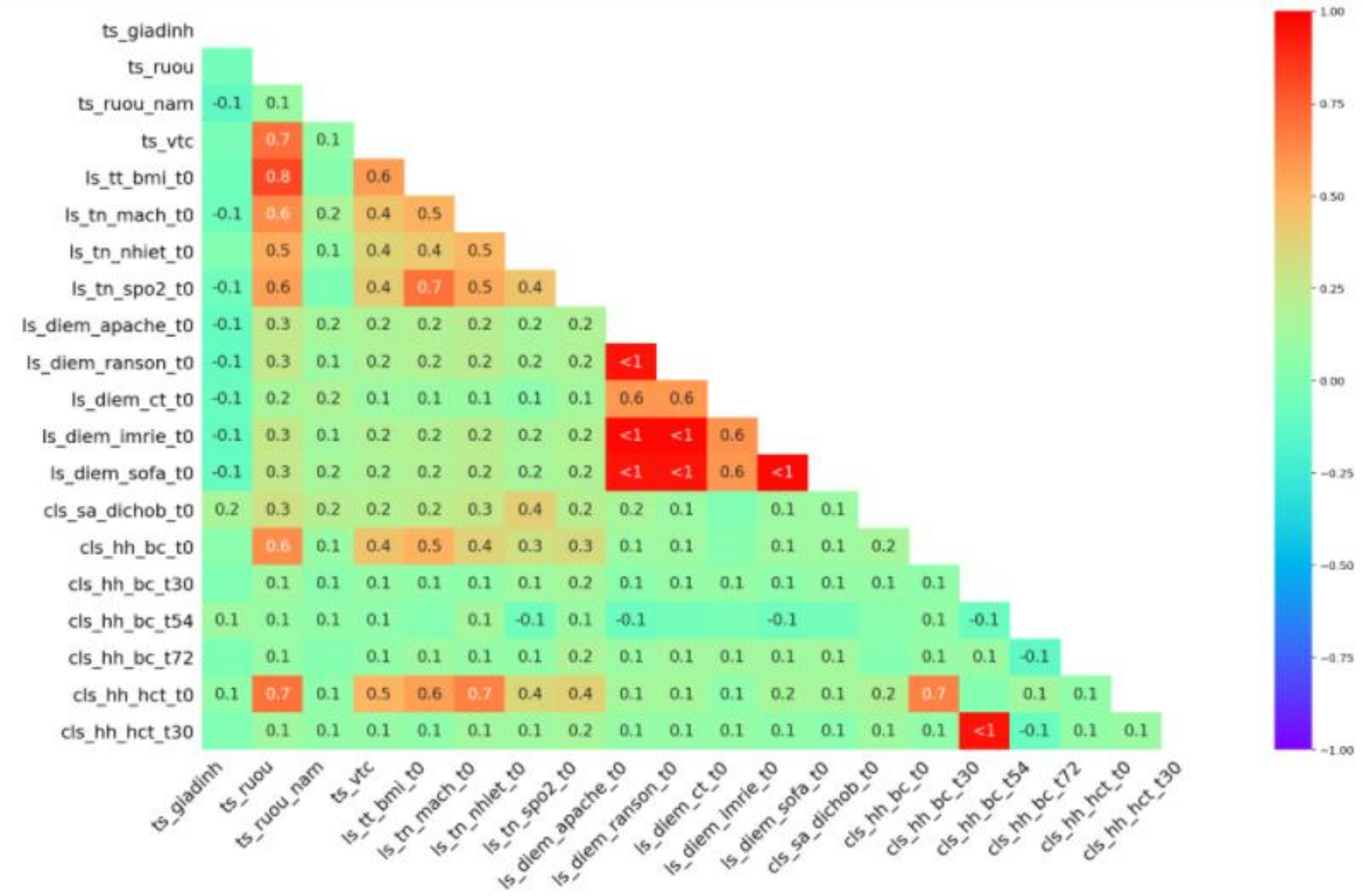


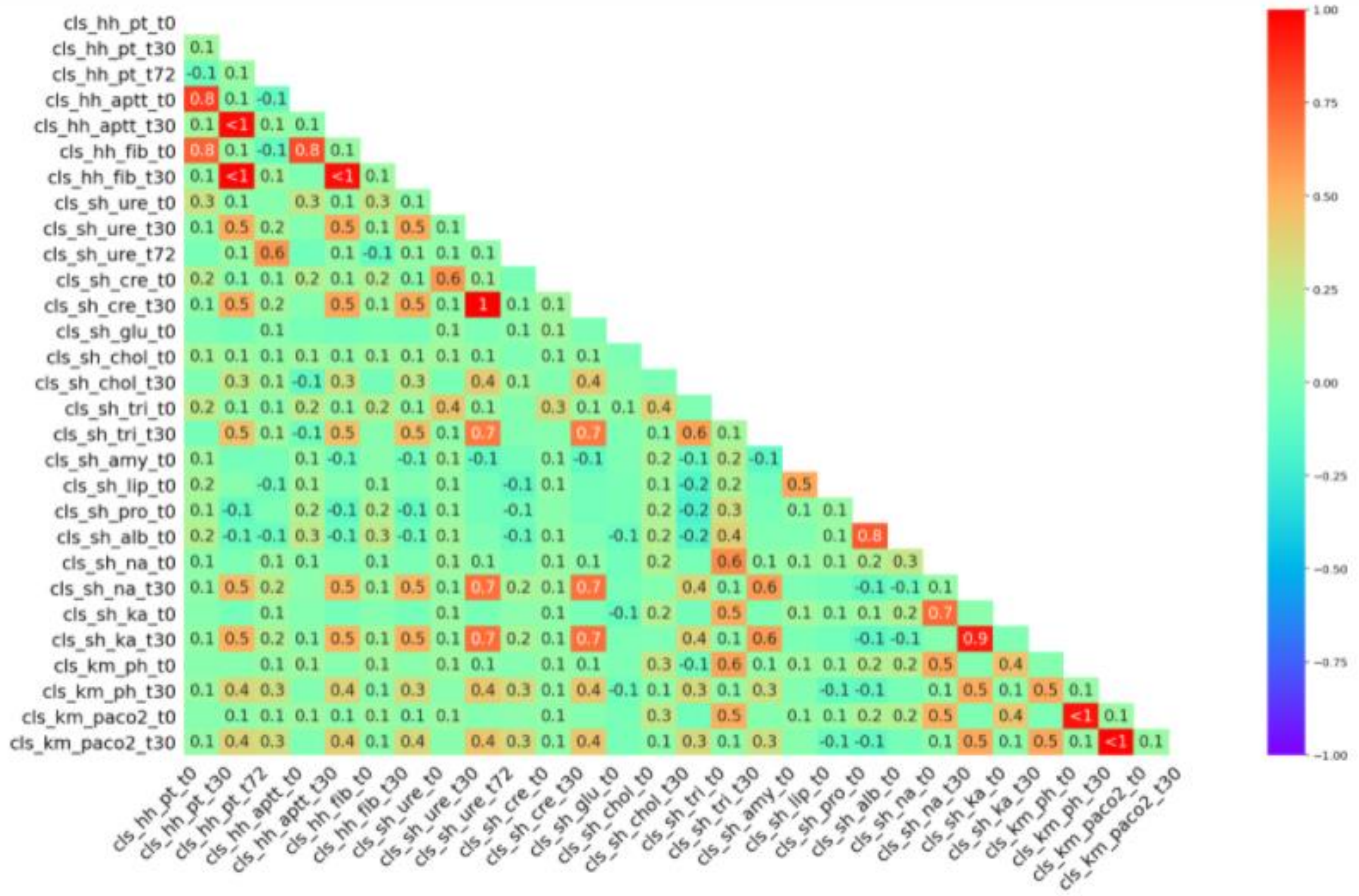


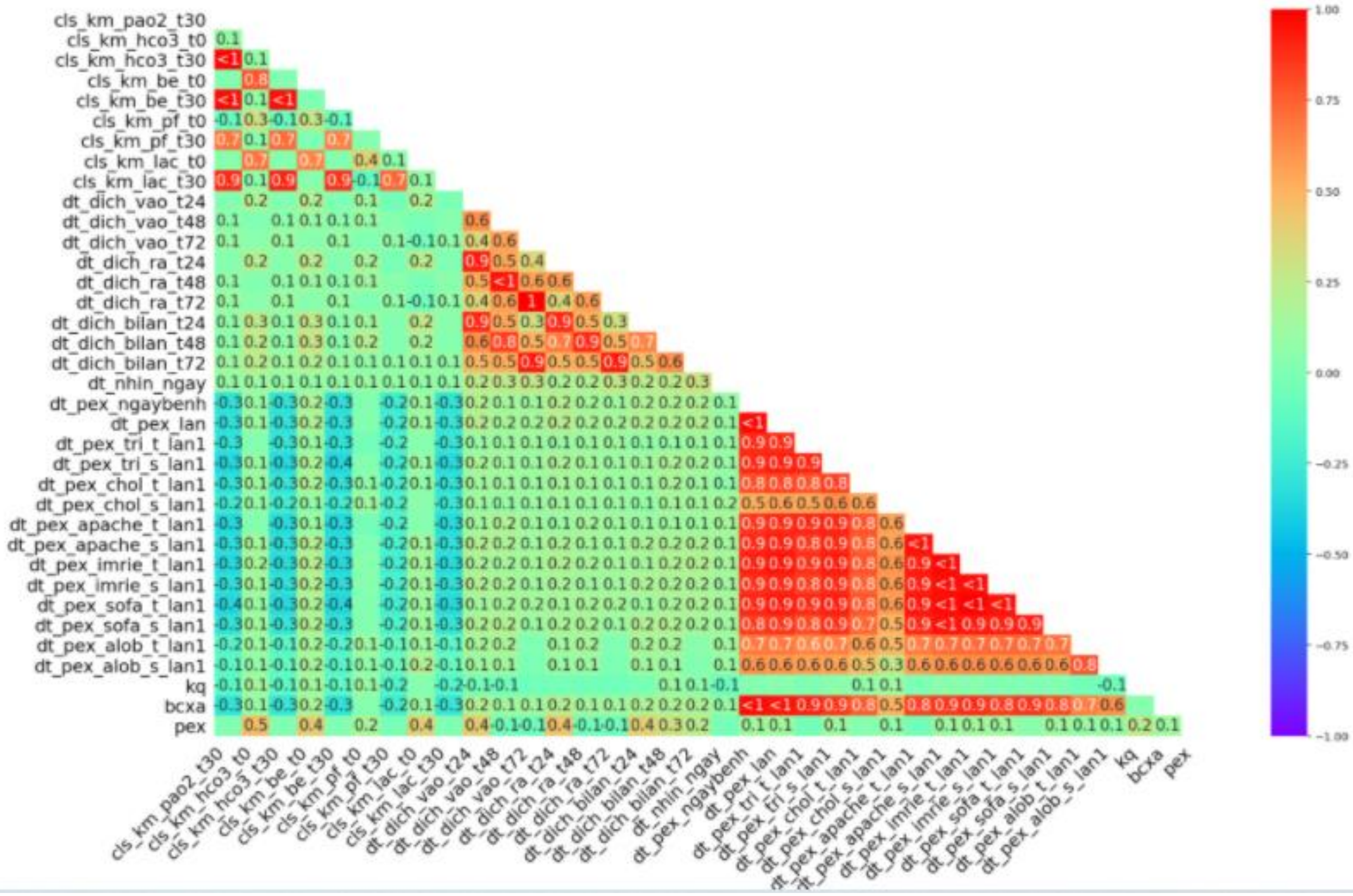


Finally, heatmaps have been developed to analyse the correlation between missing data in different variables. It enables us to identify MAR, MNAR values and also verifies the patterns observed in the charts above. The interpretation and decision of categorizing the variables as MNAR and MAR is provided in the next section.

Heatmaps of Correlation Between Missing Values of Variables:


ts_giadinh
ts_ruou
ts_ruou_nam
ts_vtc
ls_tt_bmi_t0
ls_tn_mach_t0
ls_tn_nhiet_t0
ls_tn_spo2_t0
ls_diem_apache_t0
ls_diem_ranson_t0
ls_diem_ct_t0
ls_diem_imrie_t0
ls_diem_sofa_t0
cls_sa_dichob_t0
cls_hh_bc_t0
cls_hh_bc_t30
cls_hh_bc_t54
cls_hh_bc_t72
cls_hh_hct_t0
cls_hh_hct_t30
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


cls_hh_pt_t0
cls_hh_pt_t30
cls_hh_pt_t72
cls_hh_aptt_t0
cls_hh_aptt_t30
cls_hh_fib_t0
cls_hh_fib_t30
cls_sh_ure_t0
cls_sh_ure_t30
cls_sh_ure_t72
cls_sh_cre_t0
cls_sh_cre_t30
cls_sh_glu_t0
cls_sh_chol_t0
cls_sh_chol_t30
cls_sh_tri_t0
cls_sh_tri_t30
cls_sh_amy_t0
cls_sh_lip_t0
cls_sh_pro_t0
cls_sh_alb_t0
cls_sh_na_t0
cls_sh_na_t30
cls_sh_ka_t0
cls_sh_ka_t30
cls_km_ph_t0
cls_km_ph_t30
cls_km_paco2_t0
cls_km_paco2_t30
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



Missing Data Categorization

The visualization shown in the previous section explains a lot about the nature of the missing data. Based on the analysis of visualization and domain knowledge of medical procedures, we have categorized the missing data as MAR and MNAR.

Column	Description	Category	Reason
ts_ruou	Drinking problem	MAR	There is only 1 value which is missing. Also, there is an NA value that corresponds to NO since a person without drinking problem could either fill NO or NA
ts_ruou_nam	A breakdown of drinking problem	MNAR	Only those values where patient has responded for ts_ruou as No are NA. That's because the patients without a drinking problem would be filling NA for number of years they have been drinking. Only for 2 values where ts_ruou is 'Yes', this variable is NA. These values are very few so not taking into consideration.
ts_dtd	Diabetes problem	MAR	There is only 1 value which is missing in this variable. Hence missing at random.
ts_vtc	Historical cholecystitis problem	MNAR	Although only 2 values are missing, but the pattern follows to many other variables as clear from heatmap and missing data plots.
ts_vtc_lancuoi	Last detection of cholecystitis problem	MNAR	These are dates of last detection of cholecystitis. This can be left blank due to absence of any such condition previously which is captured by ts_vtc. All the missing values correspond to answer 'NO' expect 4 which could be due to human error in filling details
non	Vomitting	MAR	As per the visualizations of missing data, missing values do not follow a pattern nor are being related to any other variable.
ls_tt_bmi_t0	Clinical symptoms : BMI	MNAR	Although only 3 values are missing, but the pattern follows that of ts_ruou i.e., drinking problem. So, patients who skipped that question also didn't fill BMI.
ls_tn_mach_t0	Clinical symptoms: Heart Rate/Pulse Per Minute	MNAR	Although data has just 5 missing values, the heatmap shows a relationship between missing values of this variable and others.
ls_tn_nhiet_t0	Body temperature	MAR	Only 7 values are missing for body temperature of patients. The visualizations shows that the missing values have very weak relationship with other variables.
ls_tn_spo2_t0	Saturation of peripheral oxygen	MNAR	The missing values of SPO2 show a high correlation with missing values of BMI. It can be seen from the patterns as well as heatmap.
ls_diem_apache_t0	apache 2 score at the points of admitting hospitals	MNAR	It can be clearly observed from the missing data patterns that these scores have highly correlated missing data. It means that either all of these were measure or none of these were measure. Thus, they are not randomly missing.
ls_diem_ranson_t0	ranson score at the points of admitting hospitals		
ls_diem_ct_t0	CTSI score at the points of admitting hospitals		

ls_diem_imrie_t0	imre score at the points of admitting hospitals		
ls_diem_sofa_t0	sofa score at the points of admitting hospitals		
cls_sa_tuy_t0	subclinical examination - (pancreas) ultrasound at the points of admitting hospitals	MNAR	For most of the values, the missing data patterns for these two variables relate to each other and to missing data of Abdominal Fluid and Bladder ultrasound exam results.
cls_sa_dichob_t0	subclinical examination - (Abdominal fluid) ultrasound at the points of admitting hospitals		
cls_sa_mat_t0	subclinical examination - (bladder) ultrasound at the points of admitting hospitals	MAR	Although the missing data results of these tests relate to previous two tests, the variables have a lot of other missing data that does not relate.
cls_ct_dichob_lan1	subclinical examination - (Abdominal fluid) computer tomography		
cls_hh_bc_t0	subclinical examination - white blood cell; t0: at the points of admitting hospitals, t6: after 6h of admitting hospitals...	MNAR	The missing values of both these variables relate to missing values of hct_t0 and hct_t30. It is because the WBC and HCT are part of a single test called CBC. Thus, if the test was not performed for some patients, all these values would be missing.
cls_hh_bc_t30	subclinical examination -		
cls_hh_bc_t54	subclinical examination -	MAR	The missing data doesn't follow a systematic pattern that relates to some other variable. Thus, the heatmap also doesn't show any correlation.
cls_hh_bc_t72	subclinical examination -	MNAR	The missing values of this variable relate to missing values of hct_t72. It is because the WBC and HCT are part of a single test called CBC. Thus, if the test was not performed for some patients, all these values would be missing.
cls_hh_hct_t0	subclinical examination - Hematocrit	MNAR	The missing values of both these variables relate to missing values of hct_t0 and hct_t30. It is because the WBC and HCT are part of a single test called CBC. Thus, if the test was not performed for some patients, all these values would be missing.
cls_hh_hct_t30	subclinical examination -		
cls_hh_hct_t72	subclinical examination -		
cls_hh_pt_t0	Prothrombin	MNAR	The missing value for these variables matches with those of APTT and Fibrinogen Tests missing values. It can be seen from the pattern as well as heat map. All these tests would been taken as a group test as they relate to each other for detecting severity. Thus, are interdependent for missing data.
cls_hh_pt_t30	subclinical examination -		
cls_hh_pt_t72	subclinical examination -	MNAR	The missing values for this variable match with the missing data of URE test at t72. Also, it has a slightly high correlation in heatmap
cls_hh_aptt_t0	APTT	MNAR	As specified in Prothrombin test, these missing values relate to each other thus absence of one is dependent on other. That's why there is a reason behind their absence and thus cannot be taken as missing at random
cls_hh_aptt_t30	subclinical examination -		
cls_hh_fib_t0	subclinical examination – Fibrinogen		
cls_hh_fib_t30	subclinical examination -		
cls_sh_ure_t0	subclinical examination – ure	MNAR	The missing values pattern matches with that of creatinine tests. Thus, there is a high correlation observed between missing values of URE at t0 and t30 and that of creatinine. Thus, it cannot be missing at random.
cls_sh_ure_t30	subclinical examination -		
cls_sh_ure_t72	subclinical examination -	MNAR	The missing values for this variable match with the missing data of APTT test at t72. Also, it has a slightly high correlation in heatmap
cls_sh_cre_t0	subclinical examination – creatinine	MNAR	As specified in Urea test, these missing values relate to each other thus absence of one is dependent on other. That's why there is a reason behind their absence and thus cannot be taken as missing at random
cls_sh_cre_t30	subclinical examination -		
cls_sh_glu_t0	subclinical examination - glucose	MAR	The missing values are occurring at random as there is no correlation in heatmap as well as no matching pattern in missing data patterns. Also, glucose is a blood sugar test that is conducted independently.
cls_sh_chol_t0	Cholesterol	MAR	The missing data is random and doesn't match with any other variable. Also, there is no strong correlation in heatmap
cls_sh_chol_t30	subclinical examination -	MNAR	The missing data is slightly like triglyceride test at t30. It can be seen from the heatmap as well with 0.6 correlation.
cls_sh_tri_t0	Triglyceride	MNAR	The missing data for triglyceride at t0 is related to missing data of natri and ph of blood air at t0. The pattern matches slightly along with a correlation between missing data of 0.6
cls_sh_tri_t30	subclinical examination -	MNAR	The missing data correlates with the missing data of UREA and Creatinine tests. There is a strong correlation between missing data as well as similar pattern.
cls_sh_amy_t0	subclinical examination – amylase	MAR	The missing data pattern for these variables are independent and do not relate to any other variable. The Heatmap also doesn't reflect any strong correlations with other variables.
cls_sh_lip_t0	subclinical examination - lipase		
cls_sh_pro_t0	subclinical examination - protein	MNAR	These two variables strongly relate to each other in terms of missing data. It reflects that once is only present when the other is present. Thus, they are not missing at random which is clearly seen in the pattern.
cls_sh_alb_t0	subclinical examination – albumin		

cls_sh_na_t0	subclinical examination – natri	MNAR	The Natri test and Potassium test have matching missing data pattern. They are also highly correlated with each other in terms of missing data. It means bot the tests are not conducted for same patients thus, there is a reason behind being missing.
cls_sh_na_t30	subclinical examination -		
cls_sh_ka_t0	subclinical examination – potassium		
cls_sh_ka_t30	subclinical examination -		
cls_km_ph_t0	subclinical examination - pH (in blood air)	MNAR	These variables are highly correlated with each other in terms of missing data. They have an approximate correlation of 1 that is perfect correlation between the missing data. It reflects that the pattern is same, and these tests are taken together.
cls_km_ph_t30	subclinical examination -		
cls_km_paco2_t0	subclinical examination - paCo2(in blood air)		
cls_km_paco2_t30	subclinical examination -		
cls_km_pao2_t0	subclinical examination - pa Oxy (in blood air)	MNAR	These variables are highly correlated with each other in terms of missing data. They have an approximate correlation of 0.9 and 1 that is almost perfect correlation between the missing data. It reflects that the pattern is same, and these tests are taken together.
cls_km_pao2_t30	subclinical examination -		
cls_km_hco3_t0	subclinical examination - HCO3- (in blood air)		
cls_km_hco3_t30			
cls_km_be_t0	BE (in blood air)		
cls_km_be_t30			
cls_km_pf_t0	p/f (paO2/%O2)	MAR	The missing data pattern is unique and doesn't follow any other variable. Also, there is no strong relationship between the missing data of variable and any other.
cls_km_pf_t30		MNAR	The pf test and lactatr test missing data aligns for t30 observations but for lac at t0 its missing values align with PAO2, HCO3 and BE test missing data. The pattern as well as correlation values are strong enough to consider them MNAR
cls_km_lac_t0	lactatr (in blood air)		
cls_km_lac_t30			
dt_dich_vao_t24	treatment - fluide intake	MNAR	All these tests are related to each other as they are the fluid test for body fluid intake and out. Thus, a missing data reflect the test was not conducted. Therefor the values are missing for every test and thus the missing data pattern match with high correlation in heatmap.
dt_dich_vao_t48	treatment - fluide intake		
dt_dich_vao_t72	treatment - fluide intake		
dt_dich_ra_t24	treatment - fluide output		
dt_dich_ra_t48	treatment - fluide output		
dt_dich_ra_t72	treatment - fluide output		
dt_dich_bilan_t24	treatment - balance fluid in and out		
dt_dich_bilan_t48	treatment - balance fluid in and out		
dt_dich_bilan_t72	treatment - balance fluid in and out		
dt_nhin_ngay	treatment - day without food intake	MAR	
dt_pex_ngaybenh	treatment - PEX treatment of which day of the diagnosis	MNAR	All these variables are observations of patients on different parameters taken before and after the PEX treatment. These values are only obtained if PEX is performed on a patient. Thus, the missing data relates to each other as no data would be present for ones not in PEX treatment. It results in high correlation of missing data between these variables as shown in heatmap
dt_pex_lan	treatment - number of PEX treatment		
dt_pex_sauvv	treatment - PEX treatment after of how many hours of the diagnosis		
dt_pex_tri_t_lan1	treatment - triglycerid before first time of PEX		
dt_pex_tri_s_lan1	treatment - triglycerid after first time of PEX		
dt_pex_chol_t_lan1	treatment - cholesterol before first time of PEX		
dt_pex_chol_s_lan1	treatment - cholesterol after first time PEX		
dt_pex_apache_t_lan1	treatment - APAche 2 score before first time PEX		
dt_pex_apache_s_lan1	treatment - APAche 2 score after first time PEX		
dt_pex_imrie_t_lan1	treatment - Imre score before first time of PEX		
dt_pex_imrie_s_lan1	treatment - Imre score after first time of PEX		
dt_pex_sofa_t_lan1	treatment - sofa score before first time of PEX		
dt_pex_sofa_s_lan1	treatment - sofa score after first time of PEX		
dt_pex_alob_t_lan1	treatment - Abdominal pressure before first time of PEX		
dt_pex_alob_s_lan1	treatment - Abdominal pressure after first time of PEX		

kq	Result - dead or alive		
bcxa	Potential complication		
pex	Patient with PEX or without PEX		

Summary

In this report, we have analysed a heavily sparse medical field data consisting of cases of Acute Pancreatitis in 165 patients who were given two kinds of treatments PEX and treatments suggested by Vietnam's Ministry of Health's guidelines in 2015. In order to analyse the data we have performed the following steps:

- Understanding the meaning of each variable in the dataset based on domain research .
- Evaluating the significance of each variable based on statistical requirement for analysis and medical relevance for analyzing pateints
- Eliminating unwanted variables based on amount of missing values and medical relevance to get a compressed dataset.
- Analyzing patterns within missing data and comparing them among all the variables in order to identify relations.
- Finally, categorized variables as Missing At Random(MAR) or Missing Not At Random(MNAR) based on patterns and relationship between the missing data and its domain significance.

APPENDIX

R CODE:

load libraries

```
library(readxl)
library(visdat)
```

```
## Warning: package 'visdat' was built under R version 4.0.5

library(naniar)

## Warning: package 'naniar' was built under R version 4.0.5

library(VIM)

## Warning: package 'VIM' was built under R version 4.0.5

## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep

library(ggplot2)
```

import dataset

```
vtc.df<-read_xlsx("APNotCleaned.xlsm", sheet =1, na = c("","NA"))
```

Filtere dataset

```
vtc.subset<- vtc.df[-c(4,5,6,7,10,11,14,18,20,21,22,23,24,28,30,37,40,41,42,44,45,47,51,52,53,54,55,57,60,61,62,63,
64,66,70,72,74,76,78,82,84,86,87,88,89,90,91,92,93,94,95,96,97,99,101,103,105,107,108,110,112,113,115,116,118,121,1
23,124,126,128,130,132,133,135,137,138,140,142,143,145,147,148,150,152,153,155,157,158,172,177,178,179,182,183,186,
187)]

n_miss(vtc.df)

## [1] 15231

pct_miss(vtc.df)

## [1] 47.58201

n_miss(vtc.subset)

## [1] 5014

pct_miss(vtc.subset)

## [1] 31.00804
```

visualize missing data

```
vis_dat(vtc.subset[1:12])

vis_dat(vtc.subset[13:24])

vis_dat(vtc.subset[25:36])

vis_dat(vtc.subset[37:48])

vis_dat(vtc.subset[49:60])

vis_dat(vtc.subset[61:72])

vis_dat(vtc.subset[73:84])

vis_dat(vtc.subset[85:98])

vis_miss(vtc.subset[1:12], show_perc = FALSE) + theme(legend.position = "right")

vis_miss(vtc.subset[13:24], show_perc = FALSE) + theme(legend.position = "right")

vis_miss(vtc.subset[25:36], show_perc = FALSE) + theme(legend.position = "right")

vis_miss(vtc.subset[37:48], show_perc = FALSE) + theme(legend.position = "right")

vis_miss(vtc.subset[49:60], show_perc = FALSE) + theme(legend.position = "right")

vis_miss(vtc.subset[61:72], show_perc = FALSE) + theme(legend.position = "right")

vis_miss(vtc.subset[73:84], show_perc = FALSE) + theme(legend.position = "right")

vis_miss(vtc.subset[85:98], show_perc = FALSE) + theme(legend.position = "right")
```

```
gg_miss_var(vtc.subset[1:12], show_pct = TRUE)

gg_miss_var(vtc.subset[13:24],show_pct = TRUE)

gg_miss_var(vtc.subset[25:36],show_pct = TRUE)

gg_miss_var(vtc.subset[37:48],show_pct = TRUE)

gg_miss_var(vtc.subset[49:60],show_pct = TRUE)

gg_miss_var(vtc.subset[61:72],show_pct = TRUE)

gg_miss_var(vtc.subset[73:84],show_pct = TRUE)

gg_miss_var(vtc.subset[85:98],show_pct = TRUE)

gg_miss_case(vtc.subset)
```

Python Code for Heatmaps

```
import pandas as pd
import numpy as np
%config InlineBackend.figure_format = 'retina'
import missingno as msno
df = pd.read_excel('Desktop/Langara/PDD Data Analytics - Langara/Fall 2021/DANA4830/Assignment1/APNotCleaned.xlsm', sheet_name='VTC-Trig-Clean', na_values=["", "NA"])
df.head()
addd=[]
for x in range(df.shape[1]):
    minus = np.array([4,5,6,7,10,11,14,18,20,21,22,23,24,28,30,37,40,41,42,44,45,47,51,52,53,54,55,57,60,61,62,63,64,66,70,72,74,76,78,82,84,86,87,88,89,90,91,92,93,94,95,96,97,99,101,103,105,107,108,110,112,113,115,116,118,121,123,124,126,128,130,132,133,135,137,138,140,142,143,145,147,148,150,152,153,155,157,158,172,177,178,179,182,183,186,187])-1
    minus = minus.tolist()
    if x not in minus:
        addd.append(x)
df = df.iloc[:,addd]
msno.heatmap(df.iloc[:,:31], cmap='rainbow')
msno.heatmap(df.iloc[:,31:60], cmap='rainbow')
msno.heatmap(df.iloc[:,60:], cmap='rainbow')
```