

✓ Mounting Google Drive (Colab specific)

```
from google.colab import drive
drive.mount('/content/drive')
```

🔗 Mounted at /content/drive

✓ Importing Required Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msno
from IPython.display import display
import re
```

```
#pd.set_option('display.max_rows', None)
#pd.set_option('display.max_columns', None)
```

✓ 1. Explore the data.

Reading the Dataset

```
df= pd.read_csv('/content/drive/MyDrive/FE/FE462.csv')
```

🔗 <ipython-input-4-b850b2b8fd5a>:1: DtypeWarning: Columns (14,15,16,17,19,20, df= pd.read_csv('/content/drive/MyDrive/FE/FE462.csv'))

1- EDA

- Lets first get to know the dataset, how many columns,
- ✓ rows, number of null values, and what are the data types

```
df.info()
```

```
>>> <class 'pandas.core.frame.DataFrame'>  
RangeIndex: 900000 entries, 0 to 899999  
Columns: 102 entries, Unnamed: 0 to 'Cystatin C'  
dtypes: float64(26), int64(2), object(74)  
memory usage: 700.4+ MB
```

```
dtypes: float64(31), int64(2), object(69)
```

```
df.memory_usage(deep=True).sum() / (1024**2)
```

```
>>> 2627.3685941696167
```

- ✓ The amount of memory used:

```
df.shape
```

```
>>> (900000, 102)
```

```
df.head()
```



```

      Unnamed: 0  RESEARCH_ID  SAMPLE_ID  COLLECTYEAR  REGN_DATE  GENDER_NAME
0

```

0	0	R015-23-48315	19152054R015-23-48315	2015.0	2015-01-01	FEMALE
1	1	R015-23-48315	19152054R015-23-48315	2015.0	2015-01-01	FEMALE
2	2	R015-23-48315	19152054R015-23-48315	2015.0	2015-01-01	FEMALE
3	3	R015-23-48315	19152054R015-23-48315	2015.0	2015-01-01	FEMALE
4	4	R015-23-48315	19152054R015-23-48315	2015.0	2015-01-01	FEMALE

```
df.tail()
```



```

      Unnamed: 0  RESEARCH_ID  SAMPLE_ID  COLLECTYEAR  REGN_DATE  GENDER
899995

```

899995	899995	R015-23-513925	172054786R015-23-513925	2020.0	2020-05-28	
899996	899996	R015-23-493532	32084900R015-23-493532	2020.0	2020-05-28	
899997	899997	R015-23-223923	32084901R015-23-223923	2020.0	2020-05-28	FE
899998	899998	R015-23-236254	4220833599R015-23-236254	2020.0	2020-05-28	
899999	899999	R015-23-236254	4220833599R015-23-236254	2020.0	2020-05-28	

```
df.sample()
```



	Unnamed: 0	RESEARCH_ID	SAMPLE_ID	COLLECTYEAR	REGN_DATE	GENDER_
102568	102568	R015-23-24152	261525172R015-23-24152	2015.0	2015-09-28	FEI

```
df.describe()
```



	Unnamed: 0	COLLECTYEAR	AGE_YEARS	AGE_DAYS	AGE_MONTHS	
count	900000.000000	838038.000000	838038.000000	838038.000000	838038.000000	90
mean	449999.500000	2017.334968	41.246880	15067.198146	502.247619	
std	259807.765474	1.575120	15.880217	5799.403257	193.315603	
min	0.000000	2015.000000	-7.000000	-2650.000000	-88.000000	
25%	224999.750000	2016.000000	30.000000	10957.000000	365.000000	
50%	449999.500000	2017.000000	39.000000	14375.000000	479.000000	
75%	674999.250000	2019.000000	52.000000	18855.000000	629.000000	
max	899999.000000	2023.000000	150.000000	54788.000000	1826.000000	16

```
df.describe(include='object')
```



	RESEARCH_ID	SAMPLE_ID	REGN_DATE	GENDER_NAME	CITY_NAME	'Thyr Stimulat Horm (TS
count	900000	900000	900000	900000	900000	545
unique	450363	776459	1967	2	34	4
top	R015-23-604	6154729R015- 23-327634	2018-09-29	FEMALE	Jeddah	
freq	174	22	3263	482928	258530	2

```
df.describe(include="all").T
```



	count	unique	top	freq	mean
Unnamed: 0	900000.0	NaN	NaN	NaN	449999.5
RESEARCH_ID	900000	450363	R015-23-604	174	NaN
SAMPLE_ID	900000	776459	6154729R015- 23-327634	22	NaN
COLLECTYEAR	838038.0	NaN	NaN	NaN	2017.334968
REGN_DATE	900000	1967	2018-09-29	3263	NaN
GENDER_NAME	900000	2	FEMALE	482928	NaN
AGE_YEARS	838038.0	NaN	NaN	NaN	41.24688
AGE_DAYS	838038.0	NaN	NaN	NaN	15067.198146
AGE_MONTHS	838038.0	NaN	NaN	NaN	502.247619
CITY_NAME	900000	34	Jeddah	258530	NaN
HEIGHT	900000.0	NaN	NaN	NaN	84.134074
WEIGHT	900000.0	NaN	NaN	NaN	123456666666706.484375
BMI	900000.0	NaN	NaN	NaN	13.285206
'Thyroid Stimulating Hormone (TSH)'	545078	4678	1.10	2821	NaN
'Uric Acid in	540028	486	4.8	11115	NaN

'Serum'	549030	400	4.0	11115	NaN
'Alanine Aminotransferase (ALT)'	580481	800	13	28059	NaN
'Ferritin In Serum'	151974	33930	< 1.00	485	NaN
'Blood Urea Nitrogen (BUN)'	361408	1925	11.0	10379	NaN
'Lymphocytes absolute count'	216.0	NaN	NaN	NaN	2.461389
'R. B. Cs / HPFs'	206	45	0-1	39	NaN
'Aspect(Urine Physical Examination)'	206	4	Clear	145	NaN
'Eosinophils absolute count'	216.0	NaN	NaN	NaN	0.193694
'Vitamin D (25 OH-Vit D -Total)'	568470	5077	11.2	2934	NaN
'C-Reactive Protein (CRP) quantitative'	11311	1110	<0.1	812	NaN
'Transferrin'	1719.0	454.0	230.0	16.0	NaN
'Height.'	0.0	NaN	NaN	NaN	NaN
'Red cell count'	216	140	5.37	4	NaN
'Basophils absolute count'	216.0	NaN	NaN	NaN	0.027269
'Crystals(Urine Microscopic Examination :)'	206	7	Absent	198	NaN
'Protein(Urine Physical Examination)'	206	7	Absent	183	NaN
'Colour(Urine Physical Examination)'	206	5	Yellow	177	NaN
'Nitrite'	206	5	Absent	201	NaN
'LDL Cholesterol'	306442	761	115	2421	NaN
'LDL / HDL'	243.0	NaN	NaN	NaN	2.852675

'24 Hour Urine Volume (263)'	7.0	NaN	NaN	NaN	2057.142857
'Hemoglobin'	219.0	NaN	NaN	NaN	13.789954
'Total Leucocytic Count'	216.0	NaN	NaN	NaN	6.62625
'Hematocrit'	219.0	NaN	NaN	NaN	40.773516
'MCV'	216.0	NaN	NaN	NaN	80.666204
'Glucose(Urine Physical Examination)'	206	4	Absent	197	NaN
'Urea in Serum'	212283	730	21	12671	NaN
'Prostatic Specific Antigen (PSA) Total'	119269	10010	0.41	294	NaN
'Testosterone (Total)'	123295	6859	4.42	281	NaN
'Alkaline Phosphatase'	288565	858	62	6103	NaN
'Total Protein in Serum'	281111.0	159.0	7.2	18238.0	NaN
'Estimated Glomerular Filtration Rate(eGFR)'	298895	86	>60	252152	NaN
'Anti CCP Abs'	6173	465	<0.5	905	NaN
' BUN/Creatinine Ratio'	275271.0	818.0	14.3	3991.0	NaN
'Blood pressure'	294405	9535	120/80	9720	NaN
'Non-HDL Cholesterol'	0.0	NaN	NaN	NaN	NaN
'Ketones'	206	3	Absent	203	NaN
'MCHC'	216.0	NaN	NaN	NaN	33.746759
'pH(Urine Physical Examination)'	206.0	NaN	NaN	NaN	5.701456
'Amorphous Elements'	206	14	Absent	178	NaN
'Blood and	206	14	Absent	178	NaN

'Haemoglobin'	206	10	Absent	176	NaN
'Epithelial Cells / HPF'	206	5	Few	149	NaN
'Casts(Urine Microscopic Examination :)'	206	3	Absent	203	NaN
'Bilirubin'	206	3	Absent	204	NaN
'Chloride in Serum'	292027.0	153.0	103.0	23388.0	NaN
'Cholesterol'	348304.0	842.0	182.0	2326.0	NaN
'T. Cholesterol/HDL'	243.0	61.0	4.2	13.0	NaN
'Urobilinogen'	206	4	Absent	140	NaN
'R.B.Cs / HPF'	5	4	0 - 1	2	NaN
'Erythrocyte Sedimentation Rate(ESR)'	144266.0	395.0	5.0	10015.0	NaN
'Glucose in Plasma (Fasting)'	523965	1181	93	11690	NaN
'Hb A1c %'	351556	463	5.2	24603	NaN
'Mean of blood glucose '	351259	868	103	12866	NaN
'Microalbuminuria (24 h urine)'	9920	3602	2	820	NaN
'Bilirubin (Total)'	321177	1187	0.40	9913	NaN
'Florescence Pattern'	2337	26	-	1203	NaN
'Lead in blood'	719	21	< 2.4	608	NaN
'Weight.'	0.0	NaN	NaN	NaN	NaN
'Monocytes absolute count'	216.0	NaN	NaN	NaN	0.505417
'Consistancy'	5	2	Semiformed	3	NaN
'Neutrophils absolute count'	216.0	NaN	NaN	NaN	3.438981
'Specific Gravity'	206	18	1.020	44	NaN
'W. B. Cs / HPF'	206	53	0-1	26	NaN

'Aspartate Aminotransferase (AST)'	574435	838	16	34178	NaN
'Calcium in Serum (Total)'	539422	287	9.5	38156	NaN
'Free T4'	390341	776	1.00	13241	NaN
'Potassium (K) in Serum'	306259	428	4.3	22460	NaN
'Albumin in Serum'	431999.0	124.0	4.4	39623.0	NaN

```
df.nunique()
```



0

Unnamed: 0	900000
RESEARCH_ID	450363
SAMPLE_ID	776459
COLLECTYEAR	9
REGN_DATE	1967
GENDER_NAME	2
AGE_YEARS	124
AGE_DAYS	30523
AGE_MONTHS	1255
CITY_NAME	34
HEIGHT	274
WEIGHT	294
BMI	2525
'Thyroid Stimulating Hormone (TSH)'	4678
'Uric Acid in Serum'	486
'Alanine Aminotransferase (ALT)'	800
'Ferritin In Serum'	33930
'Blood Urea Nitrogen (BUN)'	1925
'Lymphocytes absolute count'	142

Lymphocytes absolute count	143
'R. B. Cs / HPFs'	45
'Aspect(Urine Physical Examination)'	4
'Eosinophils absolute count'	47
'Vitamin D (25 OH-Vit D -Total)'	5077
'C-Reactive Protein (CRP) quantitative'	1110
'Transferrin'	454
'Height.'	0
'Red cell count'	140
'Basophils absolute count'	10
'Crystals(Urine Microscopic Examination :)'	7
'Protein(Urine Physical Examination)'	7
'Colour(Urine Physical Examination)'	5
'Nitrite'	5
'LDL Cholesterol'	761
'LDL / HDL'	48
'24 Hour Urine Volume (263)'	7
'Hemoglobin'	74
'Total Leucocytic Count'	188
'Hematocrit'	127
'MCV'	144
'Glucose(Urine Physical Examination)'	4
'Urea in Serum'	730
'Prostatic Specific Antigen (PSA) Total'	10010
'Testosterone (Total)'	6859
'Alkaline Phosphatase'	858
'Total Protein in Serum'	159
'Estimated Glomerular Filtration Rate(eGFR)'	86
'Anti CCP Abs'	465
' BUN/Creatinine Ratio'	818

'Blood pressure'	9535
'Non-HDL Cholesterol'	0
'Ketones'	3
'MCHC'	63
'pH(Urine Physical Examination)'	7
'Amorphous Elements'	14
'Blood and Haemoglobin'	10
'Epithelial Cells / HPF'	5
'Casts(Urine Microscopic Examination :)'	3
'Bilirubin'	3
'Chloride in Serum'	153
'Cholesterol'	842
'T. Cholesterol/HDL'	61
'Urobilinogen'	4
'R.B.Cs / HPF'	4
'Erythrocyte Sedimentation Rate(ESR)'	395
'Glucose in Plasma (Fasting)'	1181
'Hb A1c %'	463
'Mean of blood glucose '	868
'Microalbuminuria (24 h urine)'	3602
'Bilirubin (Total)'	1187
'Florescence Pattern'	26
'Lead in blood'	21
'Weight.'	0
'Monocytes absolute count'	68
'Consistancy'	2
'Neutrophils absolute count'	177
'Specific Gravity'	18
'W. B. Cs / HPF'	53
'Aspartate Aminotransferase (AST)'	838

'Calcium in Serum (Total)'	287
'Free T4'	776
'Potassium (K) in Serum'	428
'Albumin in Serum'	124
'Iron (Fe) in Serum'	17878
'CRP H.S'	6717
'Triglycerides (TG) in Serum'	1659
'Rheumatoid Factor (quantitative)'	1190
'Platelet Count'	88
'Albumin in Urine (263)'	8
'BMI'	0
'MCH'	96
'RDW'	65
'W.B.Cs / HPF'	4
'Leucocyte esterase'	7
'Concentration'	2
'Creatinine in Serum'	2362
'Sodium (Na) in Serum'	769
'Bilirubin (Direct)'	657
'Magnesium (Mg) in Serum'	60
'Titre on Hep2 cells'	11
'HDL Cholesterol'	316
'Globulin in Serum'	170
'Cystatin C'	7

dtype: int64

df.dtypes



0

Unnamed: 0

int64

RESEARCH_ID	object
SAMPLE_ID	object
COLLECTYEAR	float64
REGN_DATE	object
GENDER_NAME	object
AGE_YEARS	float64
AGE_DAYS	float64
AGE_MONTHS	float64
CITY_NAME	object
HEIGHT	int64
WEIGHT	float64
BMI	float64
'Thyroid Stimulating Hormone (TSH)'	object
'Uric Acid in Serum'	object
'Alanine Aminotransferase (ALT)'	object
'Ferritin In Serum'	object
'Blood Urea Nitrogen (BUN)'	object
'Lymphocytes absolute count'	float64
'R. B. Cs / HPFs'	object
'Aspect(Urine Physical Examination)'	object
'Eosinophils absolute count'	float64
'Vitamin D (25 OH-Vit D -Total)'	object
'C-Reactive Protein (CRP) quantitative'	object
'Transferrin'	object
'Height.'	float64
'Red cell count'	object
'Basophils absolute count'	float64
'Crystals(Urine Microscopic Examination :)'	object
'Protein(Urine Physical Examination)'	object
'Color(Urine Physical Examination)'	object

'Colour(Urine Physical Examination)'	object
'Nitrite'	object
'LDL Cholesterol'	object
'LDL / HDL'	float64
'24 Hour Urine Volume (263)'	float64
'Hemoglobin'	float64
'Total Leucocytic Count'	float64
'Hematocrit'	float64
'MCV'	float64
'Glucose(Urine Physical Examination)'	object
'Urea in Serum'	object
'Prostatic Specific Antigen (PSA) Total'	object
'Testosterone (Total)'	object
'Alkaline Phosphatase'	object
'Total Protein in Serum'	object
'Estimated Glomerular Filtration Rate(eGFR)'	object
'Anti CCP Abs'	object
' BUN/Creatinine Ratio'	object
'Blood pressure'	object
'Non-HDL Cholesterol'	float64
'Ketones'	object
'MCHC'	float64
'pH(Urine Physical Examination)'	float64
'Amorphous Elements'	object
'Blood and Haemoglobin'	object
'Epithelial Cells / HPF'	object
'Casts(Urine Microscopic Examination :)'	object
'Bilirubin'	object
'Chloride in Serum'	object
'Cholesterol'	object

'T. Cholesterol/HDL'	object
'Urobilinogen'	object
'R.B.Cs / HPF'	object
'Erythrocyte Sedimentation Rate(ESR)'	object
'Glucose in Plasma (Fasting)'	object
'Hb A1c %'	object
'Mean of blood glucose '	object
'Microalbuminuria (24 h urine)'	object
'Bilirubin (Total)'	object
'Florescence Pattern'	object
'Lead in blood'	object
'Weight.'	float64
'Monocytes absolute count'	float64
'Consistancy'	object
'Neutrophils absolute count'	float64
'Specific Gravity'	object
'W. B. Cs / HPF'	object
'Aspartate Aminotransferase (AST)'	object
'Calcium in Serum (Total)'	object
'Free T4'	object
'Potassium (K) in Serum'	object
'Albumin in Serum'	object
'Iron (Fe) in Serum'	object
'CRP H.S'	object
'Triglycerides (TG) in Serum'	object
'Rheumatoid Factor (quantitative)'	object
'Platelet Count'	object
'Albumin in Urine (263)'	float64
'BMI'	float64
'MCH'	float64

'RDW'	object
'W.B.Cs / HPF'	object
'Leucocyte esterase'	object
'Concentration'	object
'Creatinine in Serum'	object
'Sodium (Na) in Serum'	object
'Bilirubin (Direct)'	object
'Magnesium (Mg) in Serum'	object
'Titre on Hep2 cells'	object
'HDL Cholesterol'	object
'Globulin in Serum'	float64
'Cystatin C'	object

dtype: object

✓ the columns type

Unnamed: 0 int64

RESEARCH_ID object

SAMPLE_ID object

COLLECTYEAR float64

REGN_DATE object

GENDER_NAME object

AGE_YEARS float64

AGE_DAYS float64

AGE_MONTHS float64

CITY_NAME object

HEIGHT int64

WEIGHT float64

BMI float64

'Thyroid Stimulating Hormone (TSH)' object

'Uric Acid in Serum' object

'Alanine Aminotransferase (ALT)' object

'Ferritin In Serum' object

'Blood Urea Nitrogen (BUN)' object

'Lymphocytes absolute count' float64

'R. B. Cs / HPFs' object

'Aspect(Urine Physical Examination)' object

'Eosinophils absolute count' float64

'Vitamin D (25 OH-Vit D -Total)' object

'C-Reactive Protein (CRP) quantitative' object

'Transferrin' float64

'Height.' float64

'Red cell count' object

'Basophils absolute count' float64

'Crystals(Urine Microscopic Examination :)' object

'Protein(Urine Physical Examination)' object

'Colour(Urine Physical Examination)' object

'Nitrite' object

'LDL Cholesterol' object

'LDL / HDL' float64

'24 Hour Urine Volume (263)' float64

'Hemoglobin' float64

'Total Leucocytic Count' float64

'Hematocrit' float64

'MCV' float64

'Glucose(Urine Physical Examination)' object

'Urea in Serum' object

'Prostatic Specific Antigen (PSA) Total' object

'Testosterone (Total)' object

'Alkaline Phosphatase' object

'Total Protein in Serum' object

'Estimated Glomerular Filtration Rate(eGFR)' object

'Anti CCP Abs' object

' BUN/Creatinine Ratio' float64

'Blood pressure' object

'Non-HDL Cholesterol' float64

'Ketones' object

'MCHC' float64

'pH(Urine Physical Examination)' float64

'Amorphous Elements' object

'Blood and Haemoglobin' object

'Epithelial Cells / HPF' object

'Casts(Urine Microscopic Examination :)' object

'Bilirubin' object

'Chloride in Serum' object

'Cholesterol' object

'T. Cholesterol/HDL' float64

'Urobilinogen' object

'R.B.Cs / HPF' object

'Erythrocyte Sedimentation Rate(ESR)' object

'Glucose in Plasma (Fasting)' object

'Hb A1c %' object

'Mean of blood glucose ' object

'Microalbuminuria (24 h urine)' object

'Bilirubin (Total)' object
'Florescence Pattern' object
'Lead in blood' object
'Weight.' float64
'Monocytes absolute count' float64
'Consistancy' object
'Neutrophils absolute count' float64
'Specific Gravity' object
'W. B. Cs / HPF' object
'Aspartate Aminotransferase (AST)' object
'Calcium in Serum (Total)' object
'Free T4' object
'Potassium (K) in Serum' object
'Albumin in Serum' object
'Iron (Fe) in Serum' object
'CRP H.S' object
'Triglycerides (TG) in Serum' object
'Rheumatoid Factor (quantitative)' object
'Platelet Count' float64
'Albumin in Urine (263)' float64
'BMI' float64
'MCH' float64
'RDW' object
'W.B.Cs / HPF' object
'Leucocyte esterase' object
'Concentration' object
'Creatinine in Serum' object
'Sodium (Na) in Serum' object

'Bilirubin (Direct)' object

'Magnesium (Mg) in Serum' object

'Titre on Hep2 cells' object

'HDL Cholesterol' object

'Globulin in Serum' float64

'Cystatin C' float64

```
df.duplicated().sum()
```

 0


```
df['RESEARCH_ID'].duplicated().sum()
```

 449637

```
df.duplicated().sum()
```

 0

```
df.isnull().sum()
```

 0

Unnamed: 0	0
RESEARCH_ID	0
SAMPLE_ID	0
COLLECTYEAR	61962
REGN_DATE	0
GENDER_NAME	0
AGE_YEARS	61962
AGE_DAYS	61962
AGE_MONTHS	61962
CITY_NAME	0
HEIGHT	0
WEIGHT	0

BMI	0
'Thyroid Stimulating Hormone (TSH)'	354922
'Uric Acid in Serum'	350962
'Alanine Aminotransferase (ALT)'	319519
'Ferritin In Serum'	748026
'Blood Urea Nitrogen (BUN)'	538592
'Lymphocytes absolute count'	899784
'R. B. Cs / HPFs'	899794
'Aspect(Urine Physical Examination)'	899794
'Eosinophils absolute count'	899784
'Vitamin D (25 OH-Vit D -Total)'	331530
'C-Reactive Protein (CRP) quantitative'	888689
'Transferrin'	898281
'Height.'	900000
'Red cell count'	899784
'Basophils absolute count'	899784
'Crystals(Urine Microscopic Examination :)'	899794
'Protein(Urine Physical Examination)'	899794
'Colour(Urine Physical Examination)'	899794
'Nitrite'	899794
'LDL Cholesterol'	593558
'LDL / HDL'	899757
'24 Hour Urine Volume (263)'	899993
'Hemoglobin'	899781
'Total Leucocytic Count'	899784
'Hematocrit'	899781
'MCV'	899784
'Glucose(Urine Physical Examination)'	899794
'Urea in Serum'	687717
'Prostatic Specific Antigen (PSA) Total'	780721

Prostatic Specific Antigen (PSA) Total	760751
'Testosterone (Total)'	776705
'Alkaline Phosphatase'	611435
'Total Protein in Serum'	618889
'Estimated Glomerular Filtration Rate(eGFR)'	601105
'Anti CCP Abs'	893827
' BUN/Creatinine Ratio'	624729
'Blood pressure'	605595
'Non-HDL Cholesterol'	900000
'Ketones'	899794
'MCHC'	899784
'pH(Urine Physical Examination)'	899794
'Amorphous Elements'	899794
'Blood and Haemoglobin'	899794
'Epithelial Cells / HPF'	899794
'Casts(Urine Microscopic Examination :)'	899794
'Bilirubin'	899794
'Chloride in Serum'	607973
'Cholesterol'	551696
'T. Cholesterol/HDL'	899757
'Urobilinogen'	899794
'R.B.Cs / HPF'	899995
'Erythrocyte Sedimentation Rate(ESR)'	755734
'Glucose in Plasma (Fasting)'	376035
'Hb A1c %'	548444
'Mean of blood glucose '	548741
'Microalbuminuria (24 h urine)'	890080
'Bilirubin (Total)'	578823
'Florescence Pattern'	897663
'Lead in blood'	899281


'Weight.'	900000
'Monocytes absolute count'	899784
'Consistancy'	899995
'Neutrophils absolute count'	899784
'Specific Gravity'	899794
'W. B. Cs / HPF'	899794
'Aspartate Aminotransferase (AST)'	325565
'Calcium in Serum (Total)'	360578
'Free T4'	509659
'Potassium (K) in Serum'	593741
'Albumin in Serum'	468001
'Iron (Fe) in Serum'	745843
'CRP H.S'	585598
'Triglycerides (TG) in Serum'	555635
'Rheumatoid Factor (quantitative)'	889494
'Platelet Count'	899906
'Albumin in Urine (263)'	899992
'BMI'	900000
'MCH'	899784
'RDW'	899784
'W.B.Cs / HPF'	899995
'Leucocyte esterase'	899794
'Concentration'	899995
'Creatinine in Serum'	305635
'Sodium (Na) in Serum'	597210
'Bilirubin (Direct)'	580983
'Magnesium (Mg) in Serum'	882879
'Titre on Hep2 cells'	897663
'HDL Cholesterol'	594685
'Globulin in Serum'	624329

'Cystatin C'

899992

dtype: int64

```
df.isnull().sum().sum()
```

 68137483

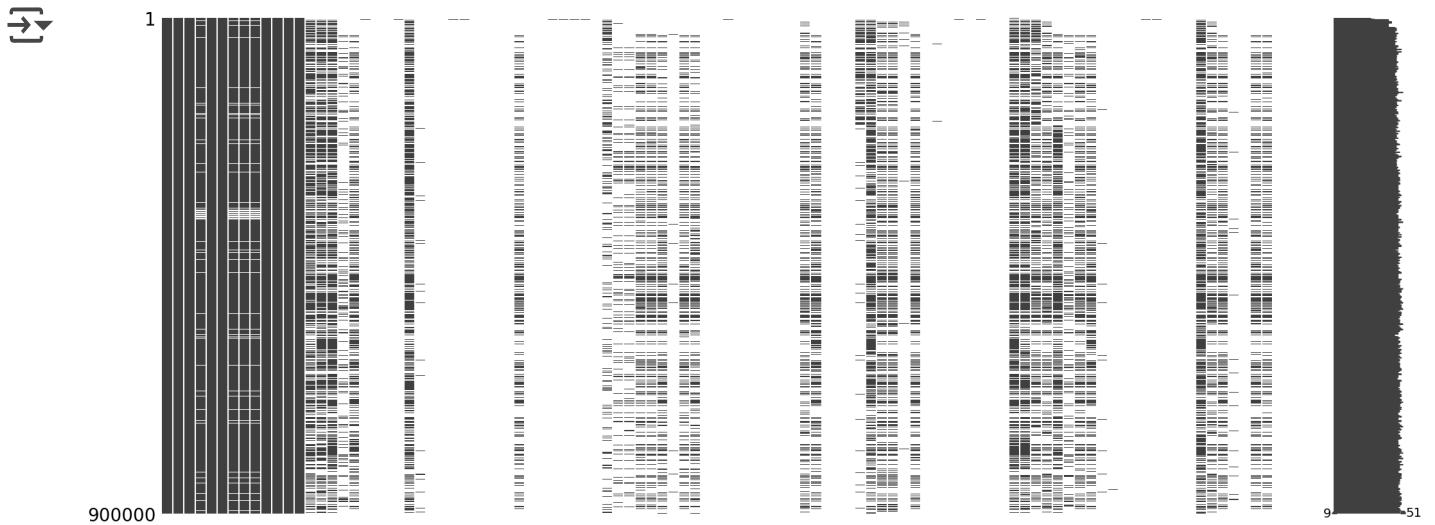
```
df.isnull().all(axis=1).sum()
```

 0

```
df.isnull().all(axis=0).sum()
```

 4


```
msno.matrix(df)
plt.show()
```



▼ -----

Missing Values Matrix Analysis

This visualization represents the missing values in the dataset.

Key Observations:

1. High Missing Values in Some Columns:

- Several columns contain a significant amount of missing data (white spaces).
- Some columns appear almost completely empty, suggesting they may not be useful.

2. Columns with Nearly Complete Data:

- A few columns (such as the one on the far right) are almost entirely filled (black).
- These columns are more reliable for analysis.

3. Patterns in Missing Data:

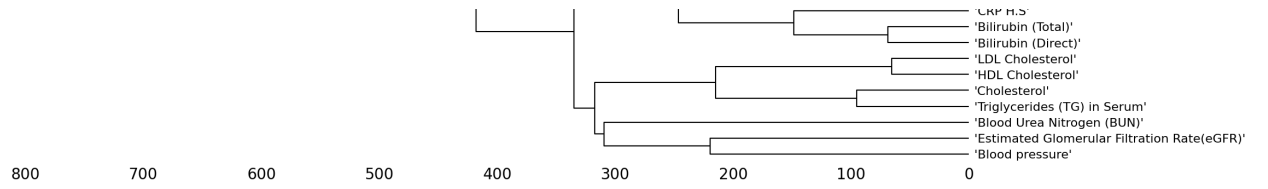
- Some missing values appear in a structured pattern across multiple rows, indicating possible systematic issues in data collection.
- Some columns show intermittent gaps, meaning certain tests or records were only available for specific samples.

4. Large Dataset Size:

- The y-axis suggests around 900,000 records, indicating a large-scale dataset.

```
msno.dendrogram(df)
plt.show()
```





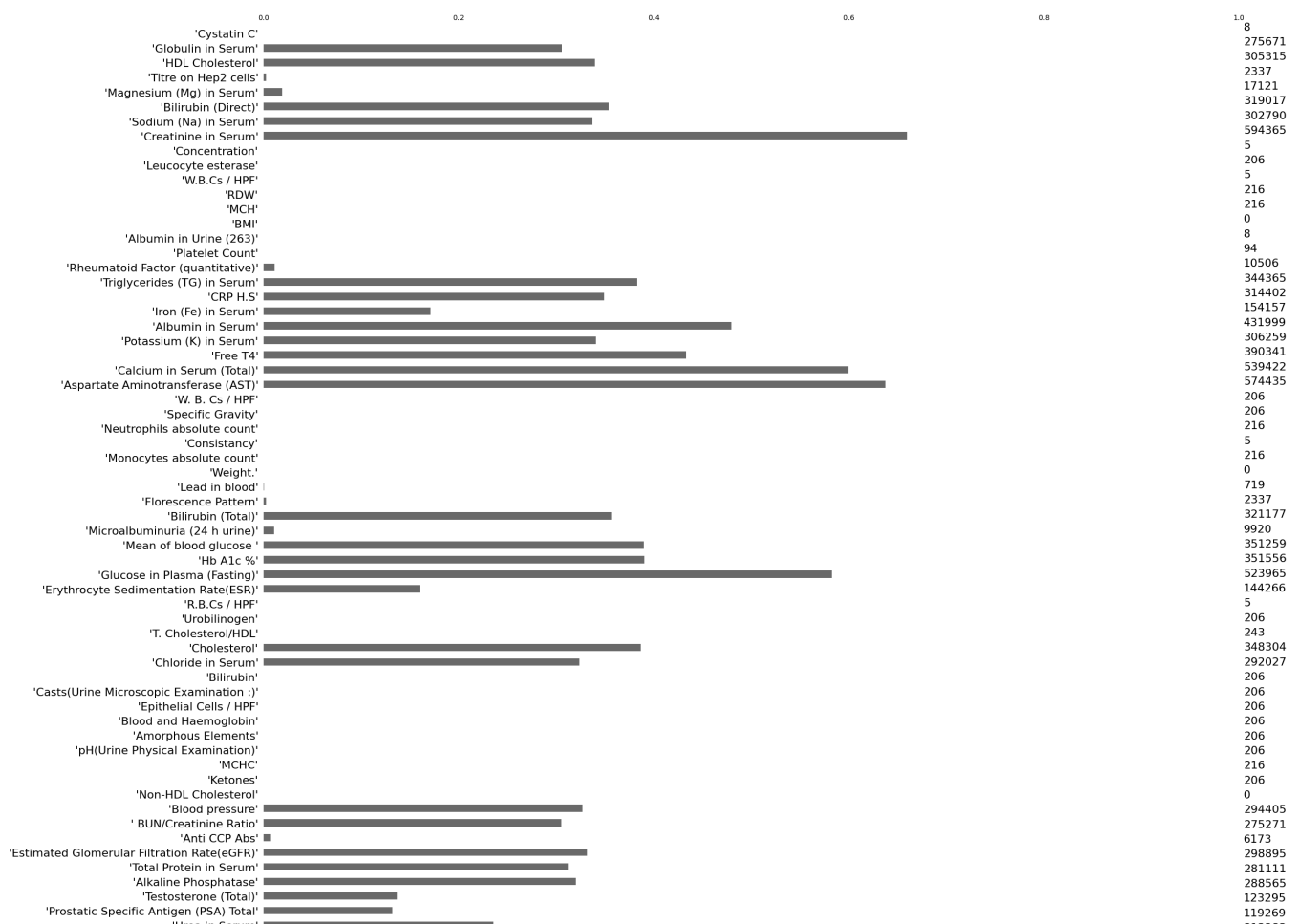
1 X-axis (Horizontal Axis):

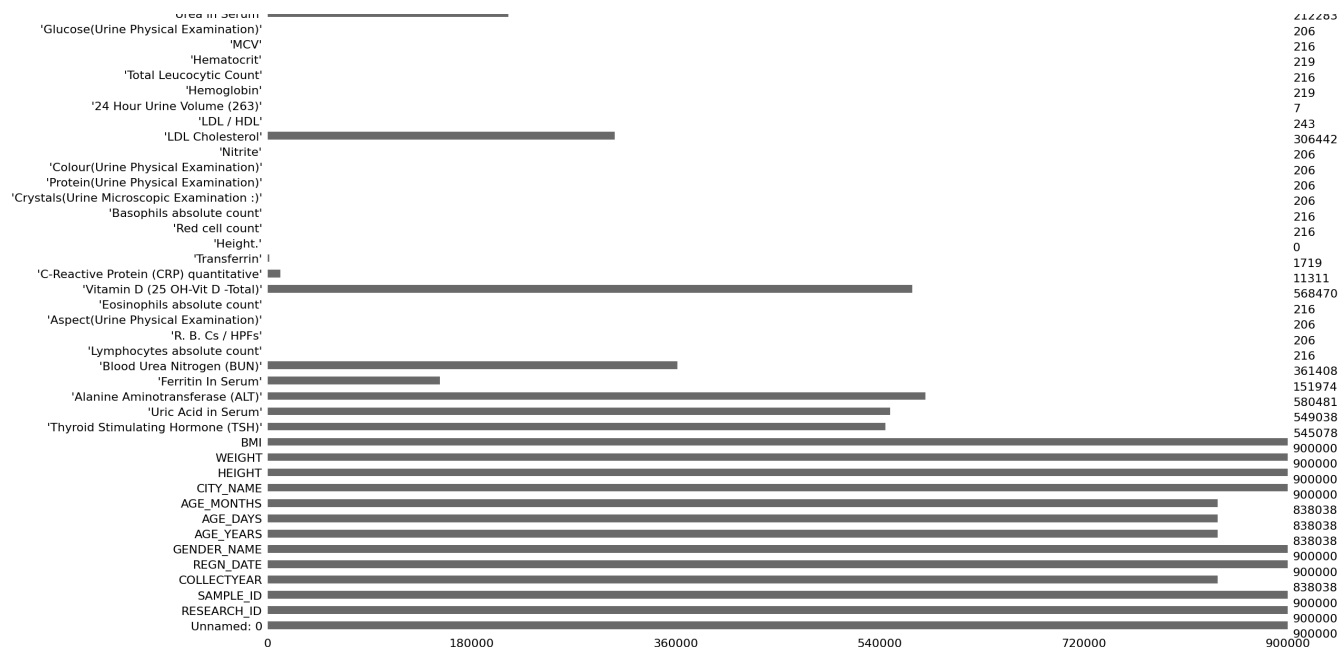
Represents the distance metric or similarity measure between variables. A larger distance indicates that the variables are more different, while a smaller distance suggests a higher similarity. **2 Y-axis (Vertical Axis):**

Displays the variables (columns) in the dataset, such as AGE_YEARS, WEIGHT, BMI, Cholesterol levels, etc. Variables are arranged in a way that clusters the most similar ones together, helping to reveal relationships in the data. **3 Branches (Clusters):**

Each branch represents a group of similar variables. The closer the connection to the bottom, the stronger the relationship between the variables. If the connection occurs at a higher level, it means the variables are less related or more distinct.

```
msno.bar(df)
plt.show()
```





1 Columns with Nearly Complete Data:

Some columns have almost all values available (close to 900,000 non-missing values).
Examples: AGE_YEARS WEIGHT HEIGHT GENDER_NAME RESEARCH_ID SAMPLE_ID These columns are highly reliable and can be used confidently in analysis.

2 Columns with Partial Missing Data:

Some columns have a moderate number of missing values, with bars positioned around the middle of the scale. Examples: Triglycerides (TG) in Serum Iron (Fe) in Serum Albumin in Serum Potassium (K) in Serum Aspartate Aminotransferase (AST) These columns can be retained but may require imputation techniques like filling missing values with the mean or median.

3 Columns with High Missing Values:

Some columns contain very few non-missing values (short bars), indicating excessive missing data. Examples: 'Cystatin C' 'Titre on Hep2 cells' 'Magnesium (Mg) in Serum' 'Leucocyte esterase' 'Basophils absolute count' These columns may be irrelevant or require removal, depending on their significance in the analysis.

```
df= df.sort_values(by='RESEARCH_ID')
```

```
print("\n".join(df.columns))
```

```

↔ Unnamed: 0
  RESEARCH_ID
  SAMPLE_ID
  COLLECTYEAR
  REGN_DATE
```

GENDER_NAME
AGE_YEARS
AGE_DAYS
AGE_MONTHS
CITY_NAME
HEIGHT
WEIGHT
BMI
'Thyroid Stimulating Hormone (TSH)'
'Uric Acid in Serum'
'Alanine Aminotransferase (ALT)'
'Ferritin In Serum'
'Blood Urea Nitrogen (BUN)'
'Lymphocytes absolute count'
'R. B. Cs / HPFs'
'Aspect(Urine Physical Examination)'
'Eosinophils absolute count'
'Vitamin D (25 OH-Vit D -Total)'
'C-Reactive Protein (CRP) quantitative'
'Transferrin'
'Height.'
'Red cell count'
'Basophils absolute count'
'Crystals(Urine Microscopic Examination :)'
'Protein(Urine Physical Examination)'
'Colour(Urine Physical Examination)'
'Nitrite'
'LDL Cholesterol'
'LDL / HDL'
'24 Hour Urine Volume (263)'
'Hemoglobin'
'Total Leucocytic Count'
'Hematocrit'
'MCV'
'Glucose(Urine Physical Examination)'
'Urea in Serum'
'Prostatic Specific Antigen (PSA) Total'
'Testosterone (Total)'
'Alkaline Phosphatase'
'Total Protein in Serum'
'Estimated Glomerular Filtration Rate(eGFR)'
'Anti CCP Abs'
' BUN/Creatinine Ratio'
'Blood pressure'
'Non-HDL Cholesterol'
'Ketones'
'MCHC'
'pH(Urine Physical Examination)'
'Amorphous Elements'
'Blood and Haemoglobin'
'Epithelial Cells / HPF'
'Casts(Urine Microscopic Examination :)'
'Bilirubin'

✓ Dataset Description

1-Identification and Demographic Information: RESEARCH_ID: A unique identifier assigned to each patient .

SAMPLE_ID: A unique identifier assigned to each biological sample collected from the patient.

COLLECTYEAR: The year in which the sample was collected.

REGN_DATE: The date when the patient or sample was registered in the system.

GENDER_NAME: The biological sex of the patient (e.g., Male, Female).

AGE_YEARS: The age of the patient in complete years.

AGE_DAYS: The exact age of the patient in days.

AGE_MONTHS: The age of the patient in months.

CITY_NAME: The city where the patient resides.

2-Physical Measurements:

HEIGHT: The patient's height measured in centimeters.

WEIGHT: The patient's weight measured in kilograms.

BMI (Body Mass Index): A measure calculated as weight (kg) divided by height squared (m²), used to assess body fat levels.

3-Thyroid and Hormonal Markers: Thyroid Stimulating Hormone (TSH): A hormone that regulates thyroid gland function.

Free T4: The free (unbound) form of thyroxine, which indicates thyroid function.

Testosterone (Total): The total testosterone level in the blood.

Prostatic Specific Antigen (PSA) Total: A marker used to assess prostate health, often for prostate cancer screening.

4-Thyroid and Hormonal Markers: Thyroid Stimulating Hormone (TSH): A hormone that regulates thyroid gland function.

Free T4: The free (unbound) form of thyroxine, which indicates thyroid function.

Testosterone (Total): The total testosterone level in the blood.

Prostatic Specific Antigen (PSA) Total: A marker used to assess prostate health, often for prostate cancer screening.

5-Liver Function and Biliary Markers: Alanine Aminotransferase (ALT): An enzyme found in the liver, with elevated levels indicating liver damage or disease.

Aspartate Aminotransferase (AST): Another liver enzyme; high levels indicate liver cell damage.

Alkaline Phosphatase: An enzyme associated with liver and bone health.

Bilirubin (Total): Measures the total amount of bilirubin in the blood, an indicator of liver function.

Bilirubin (Direct): The conjugated form of bilirubin, indicating how well the liver processes waste.

6-Lipid Profile (Blood Fat Levels): Cholesterol: The total cholesterol level in the blood, a key factor in cardiovascular health.

LDL Cholesterol: Low-density lipoprotein, known as "bad cholesterol," associated with heart disease.

HDL Cholesterol: High-density lipoprotein, known as "good cholesterol," helps remove LDL from the bloodstream.

LDL/HDL: The ratio of LDL to HDL cholesterol, used to assess cardiovascular risk.

Non-HDL Cholesterol: Total cholesterol minus HDL, representing all harmful cholesterol types.

Triglycerides (TG) in Serum: A type of fat stored in the blood; high levels are linked to heart disease.

T. Cholesterol/HDL: The ratio of total cholesterol to HDL, used for cardiovascular risk assessment.

7-Inflammatory and Immune Markers: C-Reactive Protein (CRP) quantitative: A marker of systemic inflammation.

CRP H.S: High-sensitivity CRP, a more precise measure of chronic low-grade inflammation.

Erythrocyte Sedimentation Rate (ESR): A test measuring how quickly red blood cells settle in a tube, indicating inflammation.

Rheumatoid Factor (quantitative): An antibody test used to diagnose rheumatoid arthritis.

Anti CCP Abs: Autoantibodies associated with rheumatoid arthritis.

8-Hematology (Blood Cell Analysis): Red cell count: The total number of red blood cells in a given blood volume.

Hemoglobin: The oxygen-carrying protein in red blood cells.

Hematocrit: The proportion of blood that is made up of red blood cells.

Total Leucocytic Count: The total number of white blood cells, indicating immune response.

Platelet Count: The number of platelets, important for blood clotting.

MCV (Mean Corpuscular Volume): The average size of red blood cells.

MCH (Mean Corpuscular Hemoglobin): The average hemoglobin content per red blood cell.

MCHC (Mean Corpuscular Hemoglobin Concentration): The concentration of hemoglobin within red blood cells.

RDW (Red Cell Distribution Width): A measure of variation in red blood cell size.

9-Blood Sugar and Diabetes Markers: Glucose in Plasma (Fasting): The fasting blood glucose level.

Hb A1c %: A long-term measure of blood glucose control over the past 2-3 months.

Mean of blood glucose: The estimated average blood glucose level.

10-rinalysis (Urine Examination): 24 Hour Urine Volume: The total urine volume collected over 24 hours.

Protein (Urine Physical Examination): The presence of protein in urine, which may indicate kidney disease.

Glucose (Urine Physical Examination): The presence of glucose in urine, a sign of diabetes.

Ketones: The presence of ketones in urine, indicating fat breakdown.

pH (Urine Physical Examination): The acidity or alkalinity of urine.

Specific Gravity: A measure of urine concentration.

Microalbuminuria (24 h urine): A small amount of albumin in urine, an early sign of kidney disease.

Urobilinogen: A bilirubin byproduct that can indicate liver disease.

Blood and Haemoglobin (Urine): The presence of blood in urine, which may indicate kidney disease or infection.

Nitrite: A marker of bacterial infection in urine.

Leucocyte esterase: An enzyme released by white blood cells, indicating a urinary tract infection.

R. B. Cs / HPFs: The number of red blood cells per high-power field in a urine sample.

W. B. Cs / HPF: The number of white blood cells per high-power field in a urine sample.

Crystals (Urine Microscopic Examination): The presence of crystals, which may indicate kidney stones.

Casts (Urine Microscopic Examination): Cylindrical particles in urine, often linked to kidney disease.

Aspect (Urine Physical Examination): The clarity of urine (e.g., clear, cloudy).

Colour (Urine Physical Examination): The urine color.

Amorphous Elements: Non-crystalline sediment found in urine.

Consistency: The urine's consistency or appearance.

11-Specialized Markers and Toxicology: Transferrin: A protein that transports iron.

Ferritin in Serum: The stored form of iron.

Iron (Fe) in Serum: The amount of iron in the blood.

Lead in Blood: The concentration of lead in blood, used to detect lead poisoning.

Titre on Hep2 cells: A test for autoimmune diseases.

Florescence Pattern: A laboratory result used in immunological diagnostics.

```
numeric_columns = df.select_dtypes(include=['number']).columns
ranges = df[numeric_columns].describe().loc[['min', 'max']].T
display(ranges)
```



	min	max
Unnamed: 0	0.00	8.999990e+05
COLLECTYEAR	2015.00	2.023000e+03
AGE_YEARS	-7.00	1.500000e+02
AGE_DAYS	-2650.00	5.478800e+04
AGE_MONTHS	-88.00	1.826000e+03
HEIGHT	0.00	1.681780e+05
WEIGHT	0.00	1.111110e+20
BMI	0.00	4.000000e+01
'Lymphocytes absolute count'	0.61	5.620000e+00
'Eosinophils absolute count'	0.00	5.700000e-01
'Height.'	NaN	NaN
'Basophils absolute count'	0.00	1.000000e-01
'LDL / HDL'	0.50	6.500000e+00
'24 Hour Urine Volume (263)'	200.00	6.000000e+03
'Hemoglobin'	7.60	1.860000e+01
'Total Leucocytic Count'	3.02	1.375000e+01
'Hematocrit'	24.90	5.210000e+01
'MCV'	51.30	9.340000e+01
'Non-HDL Cholesterol'	NaN	NaN
'MCHC'	28.70	3.810000e+01
'pH(Urine Physical Examination)'	5.00	8.000000e+00
'Weight.'	NaN	NaN
'Monocytes absolute count'	0.19	1.190000e+00
'Neutrophils absolute count'	0.99	8.620000e+00
'Albumin in Urine (263)'	0.05	2.002000e+01
'BMI'	NaN	NaN
'MCH'	16.50	3.280000e+01
'Globulin in Serum'	-0.30	3.740000e+01

Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.