Data Quality Report

Profile Report

Column	Data Type	Missing Values	Unique Values	Mean	Median	Std Dev
PatientID	object	0	200	nan	nan	nan
Age	int64	0	68	53.15	53.5	22.18713240193753
Gender	object	0	6	nan	nan	nan
AdmissionDate	object	0	190	nan	nan	nan
DischargeDate	object	0	4	nan	nan	nan
Diagnosis	float64	20	180	102.06957123811553	103.24515052821329	17.98083012572170 6
LabResult1	float64	0	191	5.791406770537578	5.5638012 92632213	1.7722137082903038
LabResult2	object	21	3	nan	nan	nan
Medication	int64	0	200	100.5	100.5	57.87918451 3951124
Readmitted	object	0	2	nan	nan	nan

Validation Rules

Total Anomalies Detected (Both Methods)

Count: 4

PatientID	Age	Gender	AdmissionDate	DischargeDate	Diagnosis	LabResult1	LabResult2	Medication	Readmitted	isolation_forest_anomaly	kmeans_cluster	kmeans_anomaly	combined_anomaly
34	24	Male	2023-02-03	2023-02-04	Healthy	129.41955207945455	11.53485510120990 8	Med B	No	-1	1	0	1
54	23	Female	2023-02-23	2023-02-24	COVID-19	86.3180761508 5752	11.53485510120990 8	nan	Yes	-1	1	0	1
183	33	Male	2023-07-02	2023-07-03	Hypertension	107.1937138664963	11.53485510120990 8	Med A	No	-1	1	0	1
184	76	Female	2023-07-03	2023-07-04	Asthma	123.74546 1909	11.53485510 1012	Med B	No	-1	1	0	1

PatientID	Age	Gender	AdmissionDate	DischargeDate	Diagnosis	LabResult1	LabResult2	Medication	Readmitted	isolation_forest_anomaly	kmeans_cluster	kmeans_anomaly	combined_anomaly
						34802	09908						

Anomalies Detected Only by Isolation Forest

Count: 6

PatientID	Age	Gender	AdmissionDate	DischargeDate	Diagnosis	LabResult1	LabResult2	Medication	Readmitted	isolation_forest_anomaly	kmeans_cluster	kmeans_anomaly	combined_anomaly
6	41	Female	2023-01-06	2023-01-07	Asthma	nan	4.6836860 9536731	Med C	No	-1	0	1	0
41	26	Unknown	2023-02-10	2023-02-11	Diabetes	nan	4.0668754 8778 5097	Med C	No	-1	0	1	0
67	79	Male	2023-03-08	2023-03-09	Healthy	nan	8.4158172 7438 4763	Med A	No	-1	0	1	0
164	19	Male	2023-06-13	2023-06-14	COVID-19	nan	3.4365507 9194 5335 5	Med A	No	-1	0	1	0
194	88	Female	2023-07-13	2023-07-14	Diabetes	nan	5.1646011 2103 9952	Med A	No	-1	0	1	0
199	18	Female	2023-07-18	2023-07-19	Diabetes	61.40 6031 3792 3484	7.1282281 2906 2666	nan	No	-1	0	1	0

Anomalies Detected Only by K-Means

Count: 10

PatientID	Age	Gender	AdmissionDate	DischargeDate	Diagnosis	LabResult1	LabResult2	Medication	Readmitted	isolation_forest_anomaly	kmeans_cluster	kmeans_anomaly	combined_anomaly
2	32	Male	2023-01-02	2023-01-03	Hypertension	128.6 7294 4097 6414 2	7.899 1806 4749 9765	Med A	No	1	1	0	0
49	52							nan	No	1	1	0	0

		Female	2023-02-18	2023-02-19	Healthy	135.12962 8037 6949 4	8.231 9819 5167 8605						
76	62	Female	2023-03-17	2023-03-18	Diabetes	100.6 2497 9006 8807 4	11.53 4855 1012 0990 8	Med C	No	1	1	0	0
83	28	Male	2023-03-24	2023-03-25	Hypertension	126.7 1094 1572 1783 4	8.585 6468 3162 6539	Med B	No	1	1	0	0
90	45	Male	2023-03-31	2023-04-01	Diabetes	75.11 3626 1741 8516	11.53 4855 1012 0990 8	Med C	No	1	1	0	0
103	64	Male	2023-04-13	2023-04-14	Hypertension	106.0 2340 8521 8899 8	11.53 4855 1012 0990 8	Med A	No	1	1	0	0
104	20	Male	2023-04-14	2023-04-15	Asthma	124.7 0423 5155 8077 8	11.53 4855 1012 0990 8	Med A	Yes	1	1	0	0
129	79	Female	2023-05-09	2023-05-10	Healthy	127.7 0470 4971 7140 4	11.53 4855 1012 0990 8	Med A	No	1	1	0	0
145	26	Female	2023-05-25	2023-05-26	Asthma	96.80 1893 0964 2252	8.812 9143 2287 1421	Med A	No	1	1	0	0
165	71	Unknown	2023-06-14	2023-06-15	Asthma	126.5 8566 9804 6639 4	11.53 4855 1012 0990 8	Med C	No	1	1	0	0

Imputation Suggestions
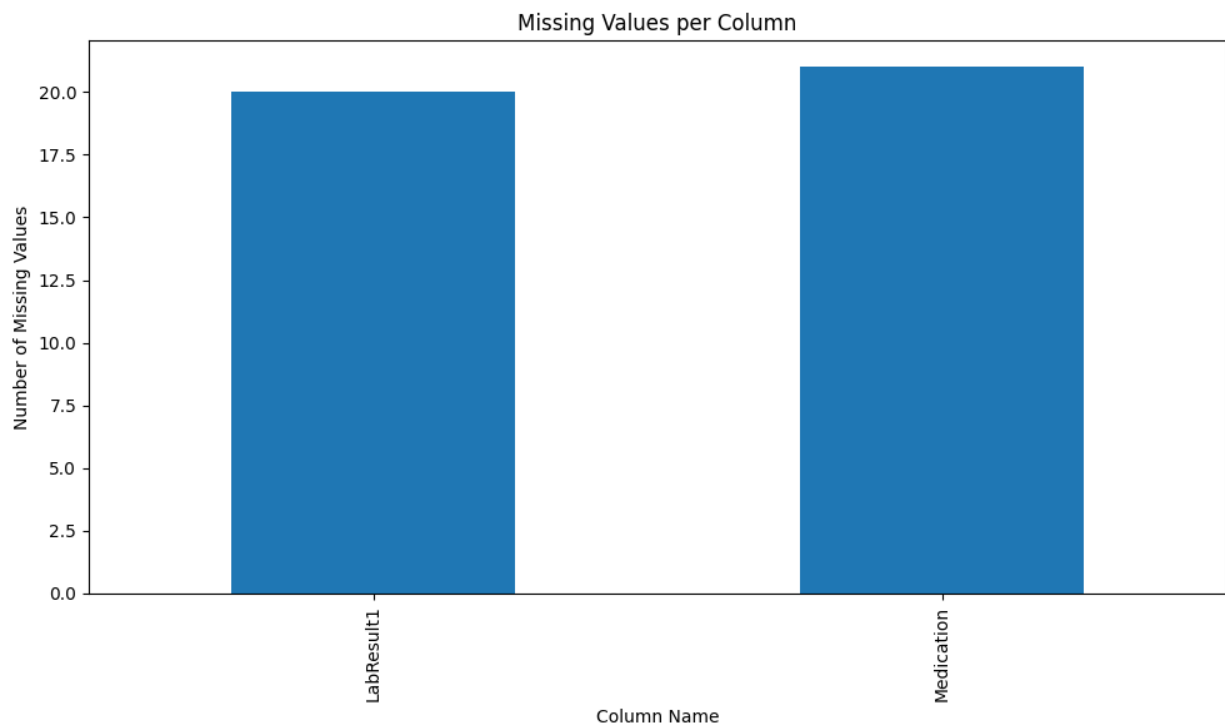
- LabResult1: Use Mean
- Medication: Use Mode

Date Validations

Date Sequencing Results

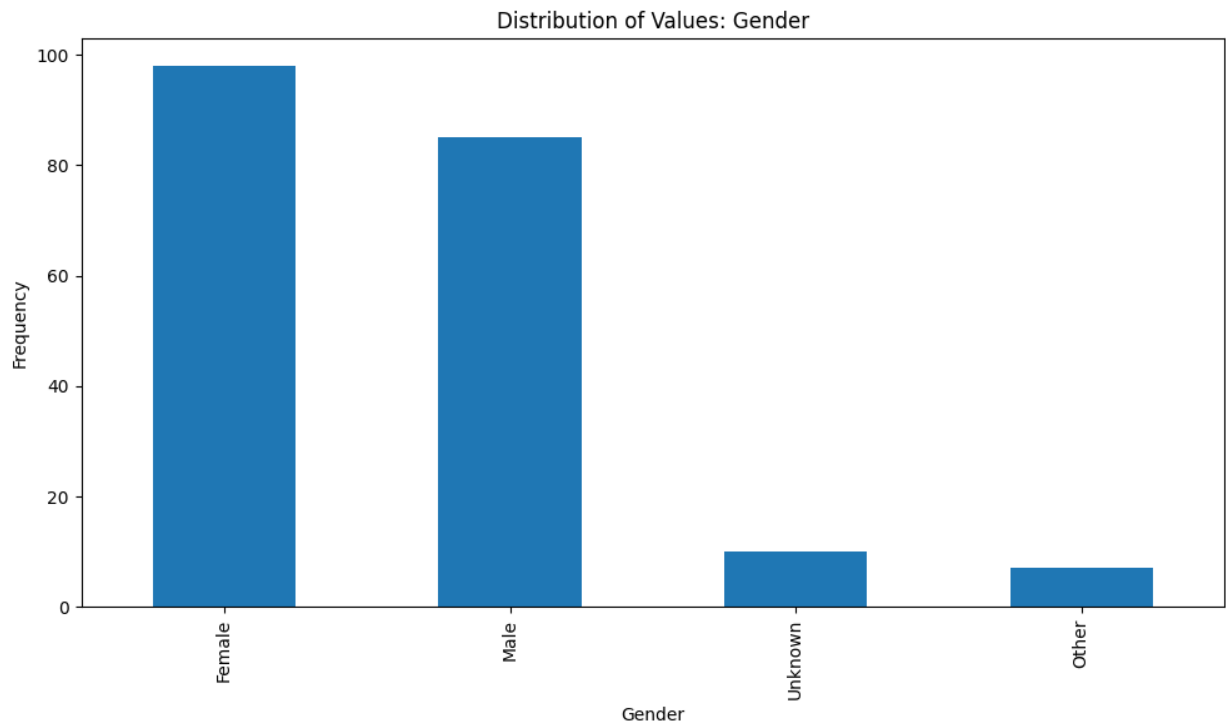Columns AdmissionDate and DischargeDate: Forward Sequences = 190, Backward Sequences = 10

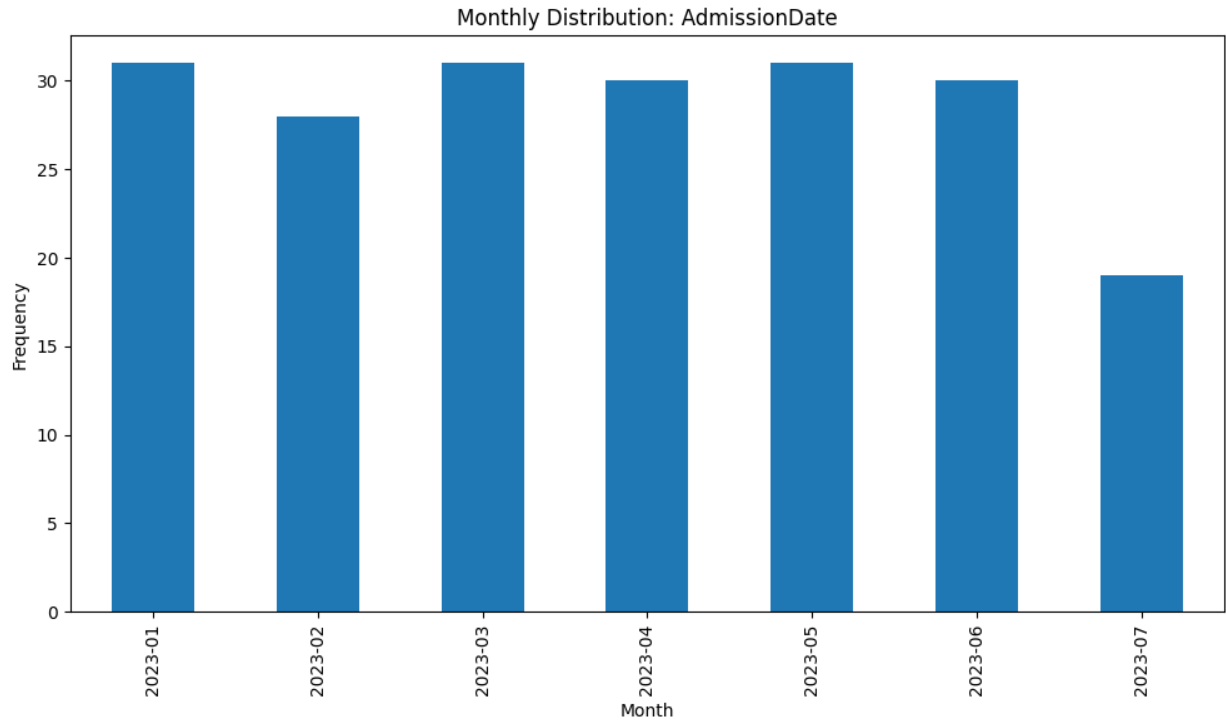Forward sequences are greater means the second date is later than the first.
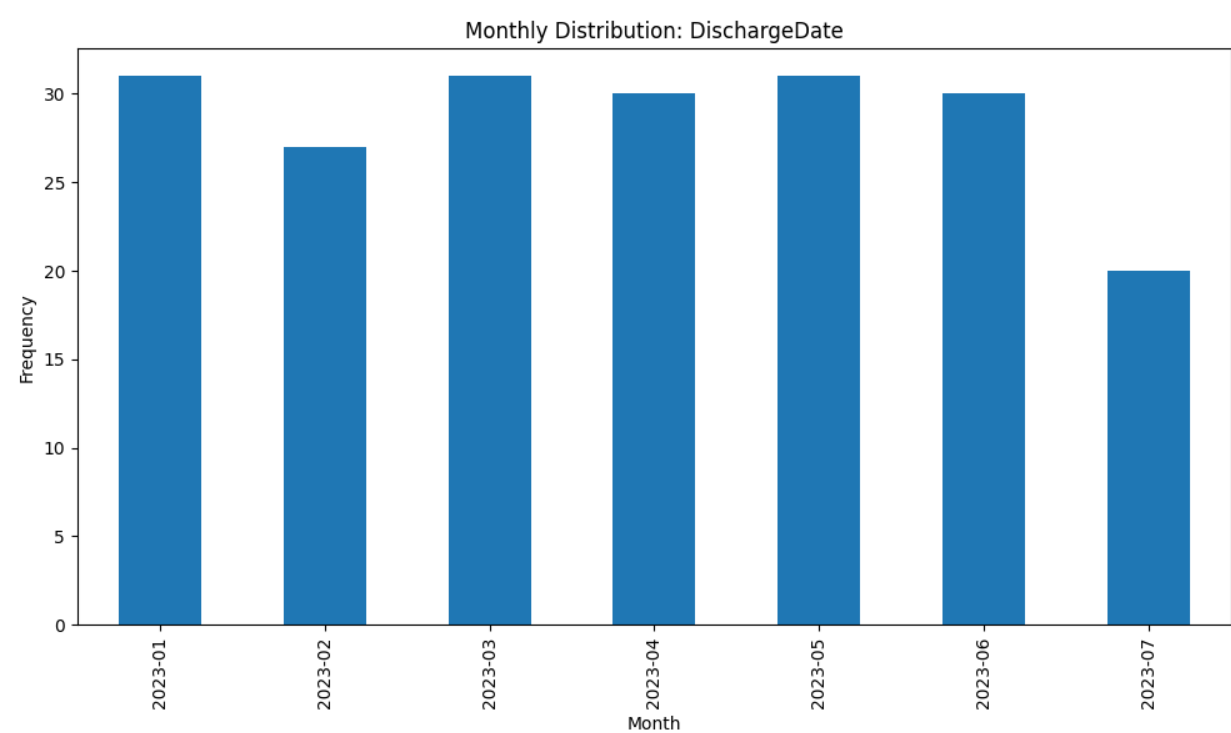
Missing Values

Missing Values per Column

Value Distributions
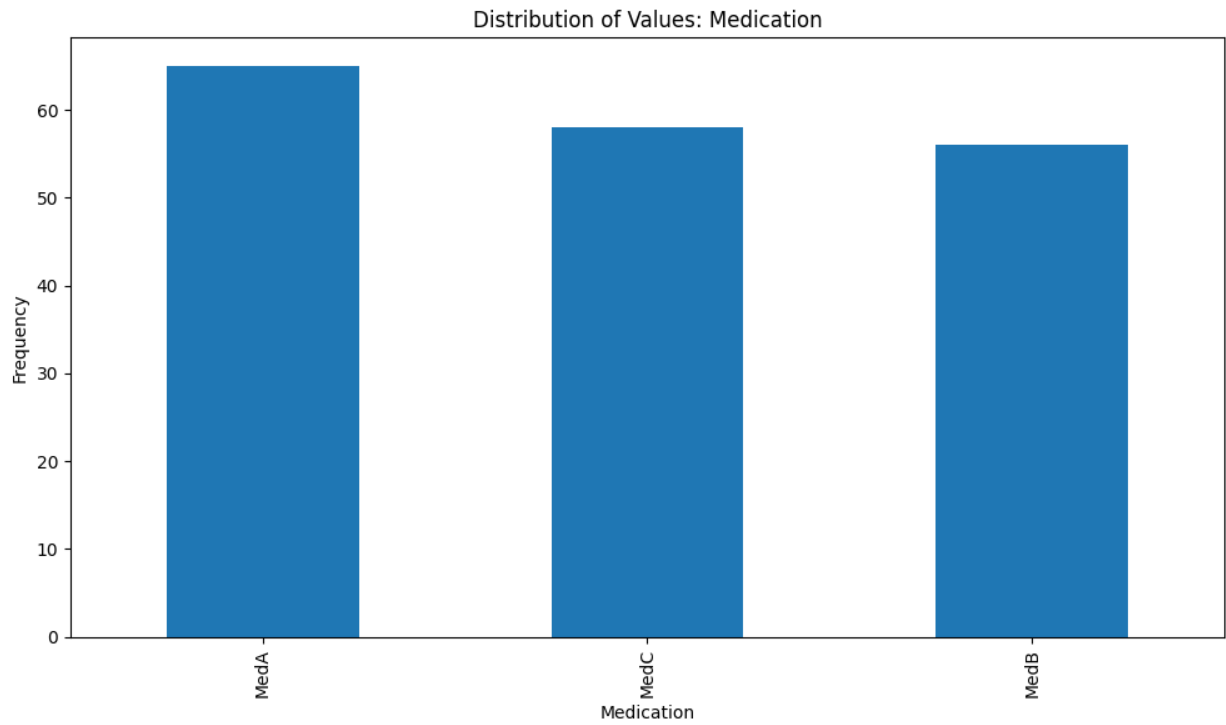
Gender

AdmissionDate

DischargeDate

Monthly Distribution: DischargeDate

Diagnosis

Distribution of Values: Diagnosis

Medication

Distribution of Values: Medication

Readmitted

Distribution of Values: Readmitted