Baza danych pogodowych

Projekt opiera się na bazie danych pogodowych dla miasta Londyn. Baza danych nie posiada braków, więc zostały one wygenerowane przy pomocy mechanizmu MCAR, MAR oraz MNAR.

Zmienne w bazie:

Zmienna	Znaczenie
cloud_cover	pomiar zachmurzenia w oktach (skala: 1-9)
sunshine	pomiar nasłonecznienia w godzinach
global_radiation	pomiar natężenia promieniowania w watach na metr kwadratowy (W/m^2)
max_temp	maksymalna zarejestrowana temperatura w stopniach Celsjusza (°C)
mean_temp	średnia temperatura w stopniach Celsjusza (°C)
min_temp	minimalna temperatura zarejestrowana w stopniach Celsjusza (°C)
precipitation	pomiar opadów w milimetrach (mm)
pressure	pomiar ciśnienia w paskalach (Pa)
snow_depth	pomiar głębokości śniegu w centymetrach (cm)
date	zarejestrowana data pomiaru

Baza danych pogodowych posiada dane od początku 2000 do końca 2020 roku. Wygląd bazy jest przedstawiony poniżej. Zmienna „date" nie będzie podlegać imputacji, jednak przy prezentacji wyników zostanie zaprezentowana.

	A	B	C	D	E	F	G	H	I	J
1	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
2	7	0,4	18	10,8	7	4,9	0	102450	0	20000101
3	7	0,7	20	11,5	7,9	5	0,2	102530	0	20000102
4	8	0	13	9,5	9,4	7,2	6	101860	0	20000103
5	5	2,9	34	11	7	4,4	0,2	101480	0	20000104
6	5	1,3	25	10,8	6,4	1,9	0,8	101420	0	20000105
7	6	0,6	20	11	8,9	7	0	101270	0	20000106
8	6	2,2	31	9,2	7,2	3,4	2	101720	0	20000107
9	4	6,4	52	7,2	7,4	5,7	0	101650	0	20000108
10	0	7,1	55	7,8	3,2	-0,7	0,2	102760	0	20000109
11	5	3,7	40	10,2	2,2	-3,3	0	103470	0	20000110
12	7	0,1	16	8,2	5,4	0,6	0	103240	0	20000111
13	7	0	14	5	8	7,8	2,8	102340	0	20000112
14	7	0	14	6,3	3,8	2,7	0,2	101540	0	20000113
15	6	4,2	45	6,7	2,9	-0,5	0,3	102330	0	20000114
16	6	0,9	25	7,2	4,6	2,6	0	103640	0	20000115
17	6	3,9	44	8,6	4,7	2,2	0	104050	0	20000116
18	5	3,8	44	8,4	4,4	0,3	0	103790	0	20000117
19	6	0	15	7,3	4,8	1,2	0,2	103540	0	20000118
20	6	0,6	24	5,3	5,4	3,6	0	103700	0	20000119
21	6	0	16	9,2	1,6	-2,2	0	103490	0	20000120
22	5	6,2	59	7,8	5	0,7	0,2	103280	0	20000121
23	5	2,1	36	7,3	5,2	2,6	1,2	102070	0	20000122
24	5	5,4	57	5,8	4	0,8	0	102420	0	20000123
25	5	6	61	3,7	3,1	0,4	0	102540	0	20000124
26	6	0,8	28	6,1	0,5	-2,7	0	102940	0	20000125
27	3	7,2	68	5,2	3,2	0,2	0	103270	0	20000126
28	0	8,1	74	10,8	0,8	-3,5	0	102970	0	20000127
29	7	0,1	20	13,5	4,5	-1,8	2,2	101480	0	20000128
30	7	3,3	49	12,9	8,4	3,4	0	100030	0	20000129
31	7	0,3	24	12,7	11,6	10,4	0	101290	0	20000130
32	7	0,3	25	10,7	11,5	10,3	0	102010	0	20000131
33	8	0	19	10,6	9,9	9,1	9,8	101520	0	20000201
34	4	6,9	74	9,3	8,2	5,8	0	102110	0	20000202

Parametry statystyk opisowych pełnego zbioru danych dla poszczególnych zmiennych (kolejno: liczba rekordów z danymi, średnia, mediana, odchylenie standardowe, współczynnik zmienności, skośność i odchylenie od średniej):

	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth
n	7671,00	7671,00	7671,00	7671,00	7671,00	7671,00	7671,00	7671,00	7671,00
x_śr	5,06	4,33	118,79	15,81	11,91	8,00	1,73	101516,70	0,02
Me	5,00	3,50	95,00	15,50	11,80	8,20	0,00	101610,00	0,00
s	2,21	3,99	88,53	6,52	5,67	5,24	3,69	1046,43	0,28
Vs	0,44	0,92	0,75	0,41	0,48	0,66	2,13	0,01	13,81
g1	-0,58	0,71	0,67	0,14	0,00	-0,17	3,79	-0,42	22,12
s(x_śr)	0,98	1,92	8,12	1,64	1,64	1,85	2,81	3,28	1,95

Macierz korelacji zmiennych:

		A	B	C	D	E	F	G	H	I
	r(i)	**cloud_cover**	**sunshine**	**global_radiation**	**max_temp**	**mean_temp**	**min_temp**	**precipitation**	**pressure**	**snow_depth**
A	**cloud_cover**	1	-0,7467	-0,4944	-0,2321	-0,1257	0,0458	0,2522	-0,2520	0,0331
B	**sunshine**	-0,7467	1	0,8400	0,4756	0,3897	0,2030	-0,2529	0,2517	-0,0491
C	**global_radiation**	-0,4944	0,8400	1	0,6954	0,6336	0,4712	-0,1826	0,1658	-0,0695
D	**max_temp**	-0,2321	0,4756	0,6954	1	0,9213	0,8152	-0,0839	0,1073	-0,1318
E	**mean_temp**	-0,1257	0,3897	0,6336	0,9213	1	0,9545	-0,0259	0,0139	-0,1480
F	**min_temp**	0,0458	0,2030	0,4712	0,8152	0,9545	1	0,0330	-0,0719	-0,1443
G	**precipitation**	0,2522	-0,2529	-0,1826	-0,0839	-0,0259	0,0330	1	-0,3630	0,0000
H	**pressure**	-0,2520	0,2517	0,1658	0,1073	0,0139	-0,0719	-0,3630	1	-0,0370
I	**snow_depth**	0,0331	-0,0491	-0,0695	-0,1318	-0,1480	-0,1443	0,0000	-0,0370	1

Najwyższe korelacje występują pomiędzy wszystkimi parametrami dotyczącymi temperatury – powyżej 0,90. Wysoka ujemna korelacja jest pomiędzy zmiennymi dotyczącymi zachmurzenia nieba oraz nasłonecznieniem w godzinach. Brak korelacji tzn. korelacja na poziomie 0 występuje pomiędzy zmiennymi dotyczącymi pomiarów opadu deszczu i śniegu.

Na zbiorze wygenerowano następujące procesy:

1. MCAR 20%
2. MAR precipitation 0: 25%; 1:10%
3. NMAR PPX cloud_cover 20%

Braki danych zostały wylosowane w programie excel. Wizualizacja i imputacja braków danych została wykonana w programie RStudio.

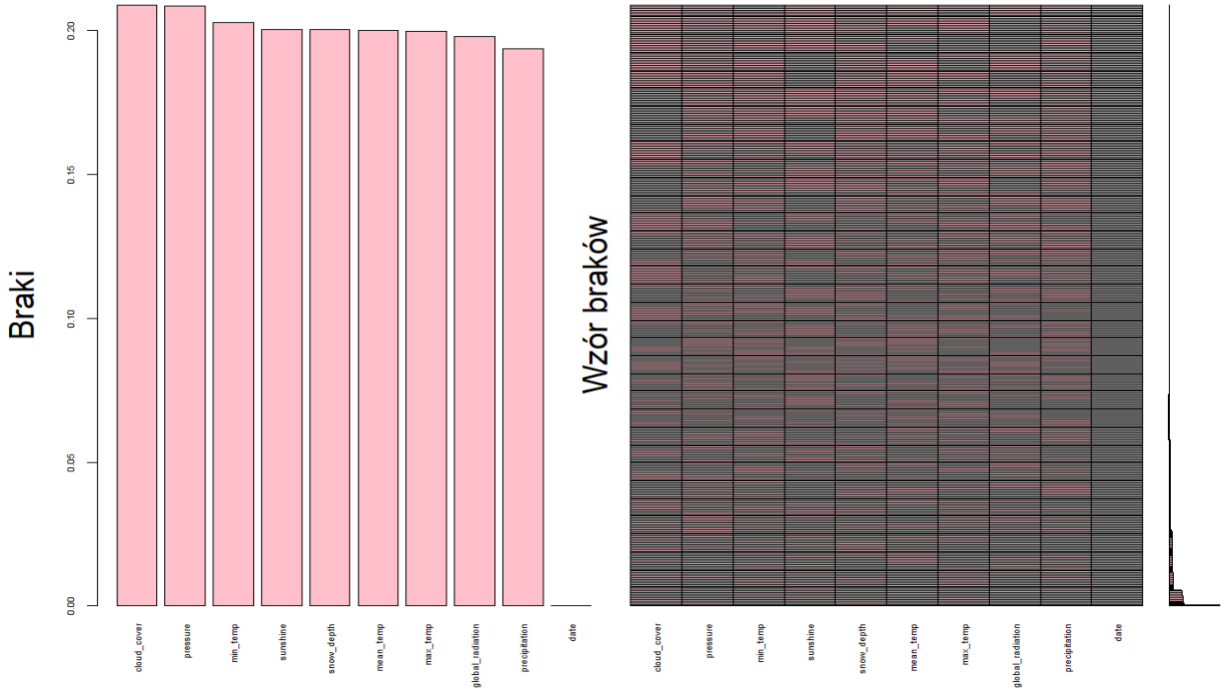
1. MCAR 20%

Mechanizm MCAR (ang. *Missing Completely at Random*) polega na całkowicie losowym występowaniu braków. Zachodzi wówczas, gdy prawdopodobieństwo wystąpienia braku danych w obrębie zmiennej Y jest niezależne od wartości tej zmiennej oraz niezależne od wartości innej zmiennej X. Braki danych zmiennej Y występują losowo wśród wszystkich obserwacji.

Prezentacja wygenerowanych braków na bazie danych:

	A	B	C	D	E	F	G	H	I	J
1	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
2	7	0	18	braki	7	5	braki	102450	braki	20000101
3	7	1	20	12	braki	5	0	102530	0	20000102
4	8	0	13	10	braki	braki	braki	101860	0	20000103
5	5	3	34	11	7	braki	0	101480	0	20000104
6	5	1	braki	11	6	braki	braki	braki	0	20000105
7	6	1	braki	11	braki	7	0	braki	0	20000106
8	6	2	31	9	7	3	2	101720	0	20000107
9	braki	6	52	7	7	6	0	101650	0	20000108
10	0	braki	55	8	braki	-1	braki	braki	0	20000109
11	5	4	40	10	braki	-3	braki	103470	braki	20000110
12	7	0	braki	braki	braki	braki	0	103240	0	20000111
13	7	braki	braki	5	braki	8	braki	braki	0	20000112
14	7	braki	14	6	4	3	0	101540	0	20000113
15	braki	4	45	7	3	-1	0	102330	0	20000114
16	6	1	25	braki	5	3	0	103640	0	20000115
17	6	4	44	9	5	2	0	104050	0	20000116
18	5	4	44	braki	4	0	0	103790	0	20000117
19	braki	braki	15	7	5	1	0	103540	0	20000118
20	6	1	24	5	5	braki	braki	103700	0	20000119

Wizualizacja braków i wzoru braków w bazie danych:



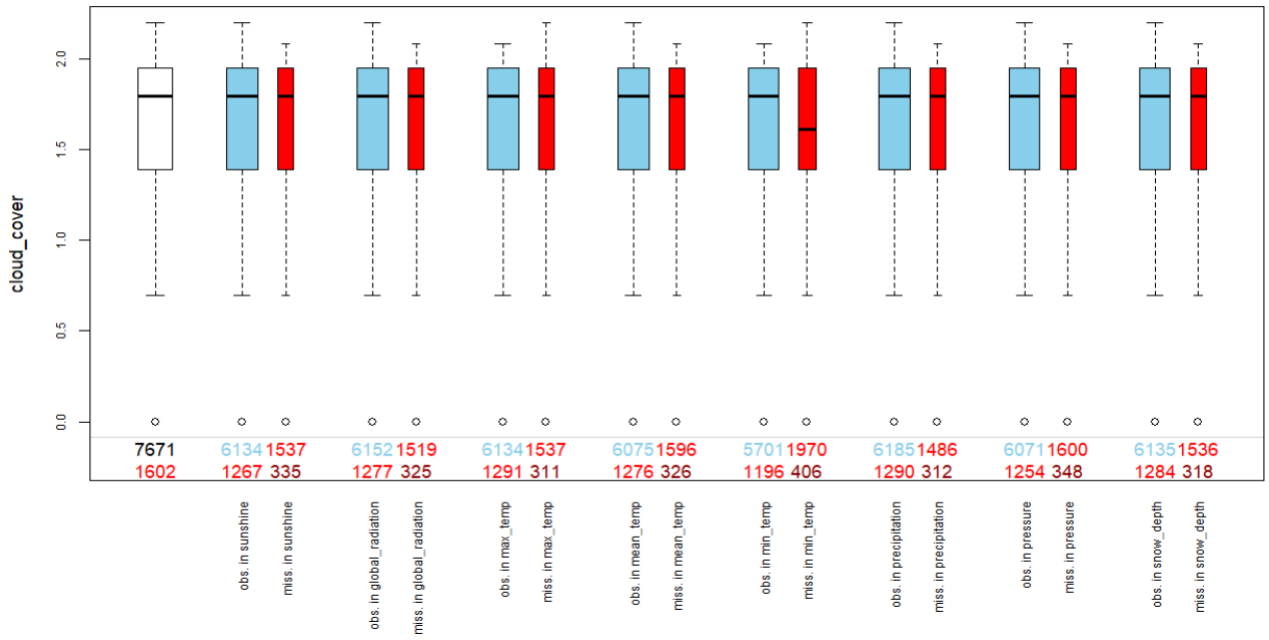
Ilości braków w zbiorze danych dla poszczególnych zmiennych:

```
cloud_cover        sunshine global_radiation        max_temp       mean_temp
       1537            1593            1491            1538            1531
   min_temp   precipitation        pressure      snow_depth            date
       1555            1495            1546            1527               0
```

Zmienne uszeregowane w kolejności od największego do najmniejszego udziału braków w zbiorze danych:

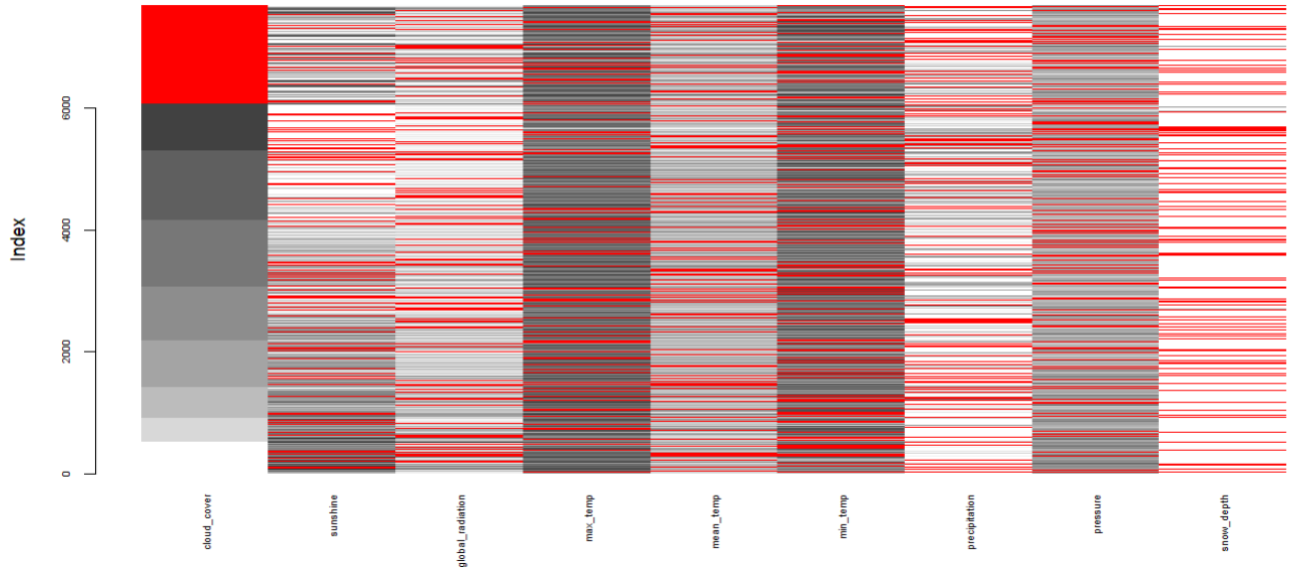
Variable	Count
sunshine	0.2076652
min_temp	0.2027115
pressure	0.2015383
max_temp	0.2004954
cloud_cover	0.2003650
mean_temp	0.1995828
snow_depth	0.1990614
precipitation	0.1948898
global_radiation	0.1943684
date	0.0000000

Wykresy pudełkowe braków – zlogarytmowane wartości obserwowane i braki wartości dla poszczególnych zmiennych:



Zgodnie z założeniem ilości braków są równomiernie rozłożone – tzn. dla każdej zmiennej występuje bardzo zbliżona ilość braków.

Rozkład braków i wartości:



Prezentacja imputacji poszczególnymi metodami:

Zmienne cloud_cover i sunshine zostały zaokrąglone do liczb całkowitych, ponieważ tak były wyrażane w zbiorze danych pierwotnych

1) k najbliższych sąsiadów (z odległością Gowera)

MCAR1<-kNN(MCAR, numFun = weightedMean, weightDist=TRUE)

MCARkNN<-MCAR1[,-c(11:20)]

	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
1	7	0	18.00000	11.028396	7.000000	4.9000000	1.56217439	102450.00	0.0000000	20000101
2	7	1	20.00000	11.500000	7.222965	5.0000000	0.20000000	102530.00	0.0000000	20000102
3	8	0	13.00000	9.500000	6.866766	5.7042532	2.23810817	101860.00	0.0000000	20000103
4	5	3	34.00000	11.000000	7.000000	3.2787919	0.20000000	101480.00	0.0000000	20000104
5	5	1	33.34387	10.800000	6.400000	4.3634783	1.52385115	101157.48	0.0000000	20000105
6	6	1	49.84065	11.000000	8.723905	7.0000000	0.00000000	100265.86	0.0000000	20000106
7	6	2	31.00000	9.200000	7.200000	3.4000000	2.00000000	101720.00	0.0000000	20000107
8	4	6	52.00000	7.200000	7.400000	5.7000000	0.00000000	101650.00	0.0000000	20000108
9	0	4	55.00000	7.800000	4.403329	-0.7000000	0.23863216	102242.61	0.0000000	20000109
10	5	4	40.00000	10.200000	4.934899	-3.3000000	0.15882608	103470.00	0.0000000	20000110

2) Regresja

MCAR_REG <- regressionImp(cloud_cover+sunshine+global_radiation+min_temp+mean_temp+

max_temp+precipitation+pressure+snow_depth~date,data=MCAR)

MCAR_REG<-MCAR_REG[,-c(11:19)]

	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
1	7	0	18.0000	15.28939	7.00000	4.900000	1.845870	102450.0	0.01943003	20000101
2	7	1	20.0000	11.50000	11.56165	5.000000	0.200000	102530.0	0.00000000	20000102
3	8	0	13.0000	9.50000	11.56165	7.885472	1.845869	101860.0	0.00000000	20000103
4	5	3	34.0000	11.00000	7.00000	7.885473	0.200000	101480.0	0.00000000	20000104
5	5	1	119.1253	10.80000	6.40000	7.885475	1.845867	101526.7	0.00000000	20000105
6	6	1	119.1253	11.00000	11.56166	7.000000	0.000000	101526.7	0.00000000	20000106
7	6	2	31.0000	9.20000	7.20000	3.400000	2.000000	101720.0	0.00000000	20000107
8	5	6	52.0000	7.20000	7.40000	5.700000	0.000000	101650.0	0.00000000	20000108
9	0	5	55.0000	7.80000	11.56167	-0.700000	1.845863	101526.7	0.00000000	20000109
10	5	4	40.0000	10.20000	11.56168	-3.300000	1.845862	103470.0	0.01943005	20000110

3) Random Forest

MCAR_RandF<-rangerImpute(cloud_cover+sunshine+global_radiation+min_temp+mean_temp+

max_temp+precipitation+pressure+snow_depth~date,data=MCAR)

MCAR_RandF<-MCAR_RandF[,-c(11:19)]

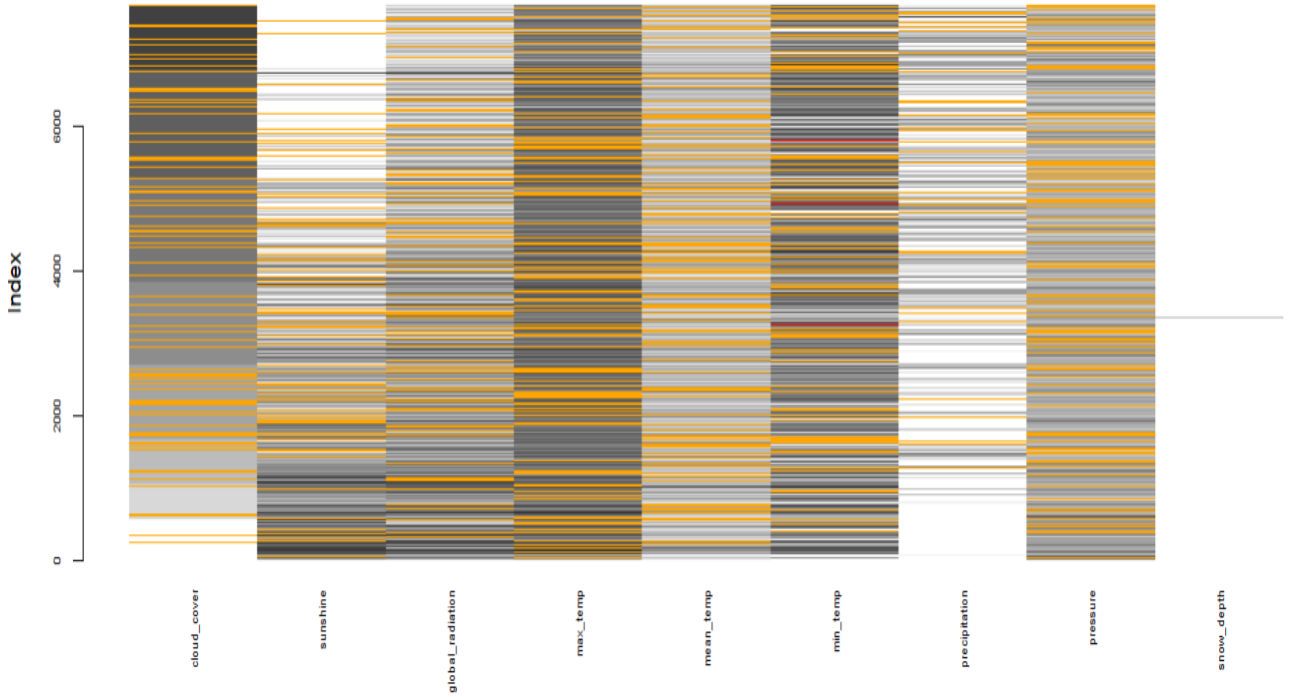
	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
1	7	0	18.00000	10.744810	7.000000	4.9000000	0.48575667	102450.0	0	20000101
2	7	1	20.00000	11.500000	6.920067	5.0000000	0.20000000	102530.0	0	20000102
3	8	0	13.00000	9.500000	6.920067	5.0537667	0.48575667	101860.0	0	20000103
4	5	3	34.00000	11.000000	7.000000	5.0816467	0.20000000	101480.0	0	20000104
5	5	1	30.38047	10.800000	6.400000	5.1899500	0.48443667	101795.7	0	20000105
6	6	1	35.63877	11.000000	6.892553	7.0000000	0.00000000	101806.4	0	20000106
7	6	2	31.00000	9.200000	7.200000	3.4000000	2.00000000	101720.0	0	20000107
8	5	6	52.00000	7.200000	7.400000	5.7000000	0.00000000	101650.0	0	20000108
9	0	4	55.00000	7.800000	6.994673	-0.7000000	0.30297333	102131.2	0	20000109
10	5	4	40.00000	10.200000	6.756370	-3.3000000	0.15721000	103470.0	0	20000110

Zestawienie statystyk opisowych zbiorów danych po wykonanych imputacjach do pełnego zbioru danych oraz zbioru danych z wygenerowanymi brakami:

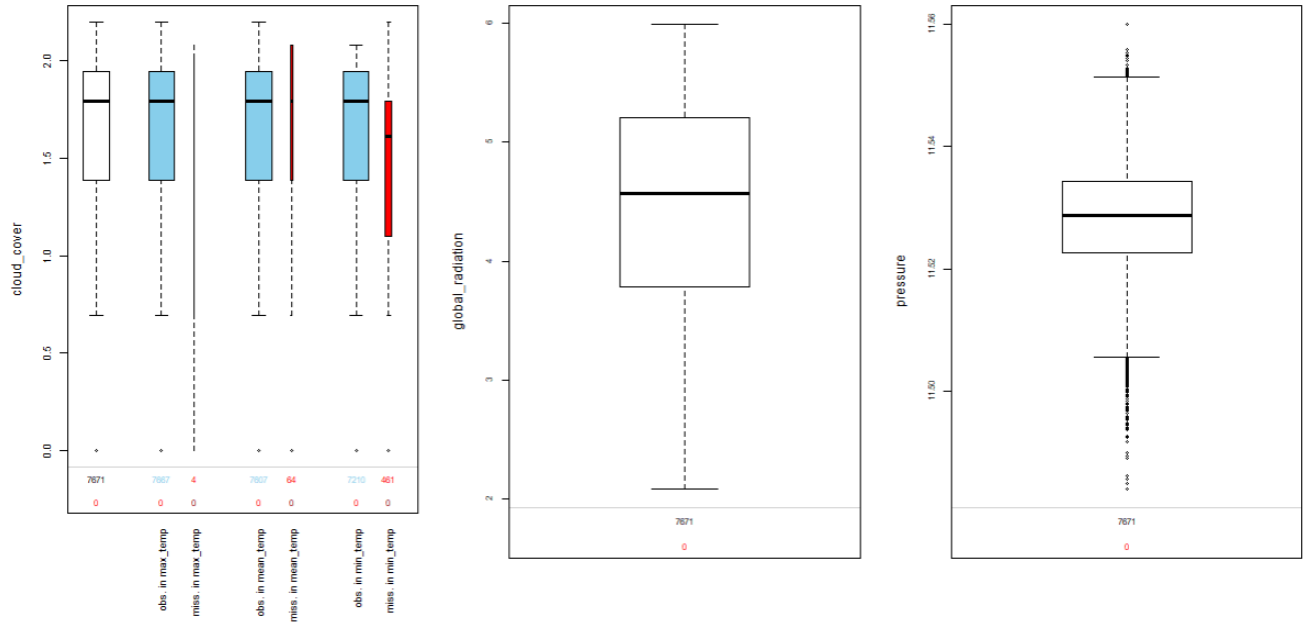
	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth
	Pełen zbiór danych								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	5,06	4,33	118,79	15,81	11,91	8	1,73	101516,7	0,02
Me	5	3,5	95	15,5	11,8	8,2	0	101610	0
s	2,21	3,99	88,53	6,52	5,67	5,24	3,69	1046,43	0,28
Vs	0,44	0,92	0,75	0,41	0,48	0,66	2,13	0,01	13,81
g1	-0,58	0,71	0,67	0,14	0	-0,17	3,79	-0,42	22,12
s(x_śr)	0,98	1,92	8,12	1,64	1,64	1,85	2,81	3,28	1,95
	Zbiór danych z brakami wywołanymi mechanizmem MCAR								
n	6069	6134	6152	6138	6137	6116	6185	6071	6135
x_śr	5,06	4,34	119,03	15,81	11,91	8,02	1,76	101515,39	0,02
Me	5	3,5	95	15,5	11,8	8,3	0,1	101600	0
s	2,2	3,99	89,05	6,52	5,71	5,24	3,77	1045,94	0,25
Vs	0,44	0,92	0,75	0,41	0,48	0,65	2,14	0,01	12,67
g1	-0,58	0,72	0,67	0,14	0	-0,17	3,86	-0,4	17,12
s(x_śr)	0,98	1,92	8,16	1,64	1,65	1,85	2,84	3,28	1,78
	Imputacja metodą k najbliższych sąsiadów (z odległością Gowera)								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	5,07	4,29	118,47	15,83	11,94	8,02	1,76	101526,17	0,02
Me	5	4	96	15,6	11,8	8,2	0,2	101592,87	0
s	2,08	3,87	86	6,32	5,55	5,09	3,51	973,64	0,24
Vs	0,41	0,9	0,73	0,4	0,46	0,63	2	0,01	11,64
g1	-0,6	0,72	0,67	0,13	-0,01	-0,16	3,93	-0,42	16,92
s(x_śr)	0,93	1,87	7,9	1,59	1,61	1,8	2,65	3,06	1,66
	Imputacja regresją								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	5,05	4,27	119,03	15,81	11,91	8,02	1,76	101515,37	0,02
Me	5	4	118,97	15,76	11,9	8,03	0,4	101518,71	0
s	1,96	3,59	79,75	5,83	5,11	4,68	3,38	930,48	0,22
Vs	0,39	0,84	0,67	0,37	0,43	0,58	1,92	0,01	11,33
g1	-0,64	0,83	0,75	0,15	0	-0,19	4,3	-0,45	19,14
s(x_śr)	0,87	1,74	7,31	1,47	1,48	1,65	2,55	2,92	1,59
	Imputacja metodą Random Forest – lasy losowe								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	5,06	4,32	118,89	15,8	11,9	8	1,74	101518,57	0,02
Me	5	4	97,35	15,6	11,8	8,2	0,2	101600	0
s	2,09	3,83	87,03	6,46	5,63	5,17	3,54	1011,84	0,24
Vs	0,41	0,89	0,73	0,41	0,47	0,65	2,03	0,01	12,47
g1	-0,6	0,72	0,64	0,13	0	-0,16	3,93	-0,39	17,38
s(x_śr)	0,93	1,84	7,98	1,62	1,63	1,83	2,68	3,18	1,72

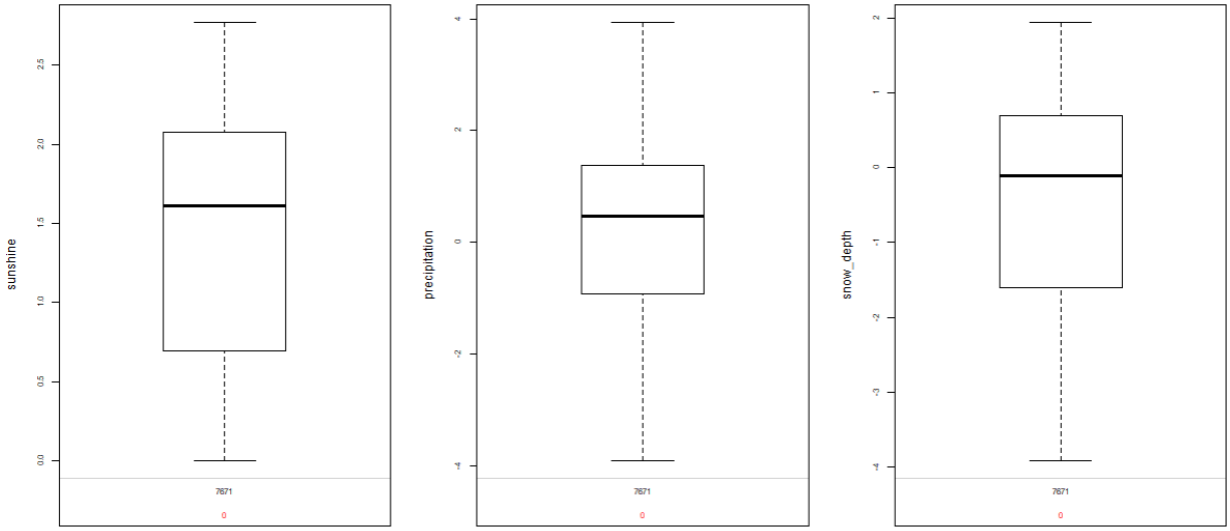
Na podstawie powyższej tabeli można zauważyć, że wszystkie metody imputacji działają poprawnie, ponieważ nie występują żadne braki danych. Najwięcej brakujących danych zostało wylosowanych dla zmiennych cloud_cover i preassure. Porównując średnią, medianę, odchylenie standardowe, współczynnika zmienności oraz odchylenia od średniej wszystkich zmiennych najlepiej zadziałała imputacja za pomocą lasów losowych. W przypadku skośności najlepsze wyniki daje metoda k najbliższych sąsiadów, jednak wyniki są porównywalne z metodą lasów losowych. Imputacja za pomocą regresji zdecydowanie jest tutaj nietrafiona – wyniki zdecydowanie odbiegają od wartości opisujących pełen zbiór danych.

Wizualizacja rozkładu braków po imputacji metodą k najbliższych sąsiadów (z odległością Gowera):



Wykresy pudełkowe po imputacji metodą k najbliższych sąsiadów (z odległością Gowera):





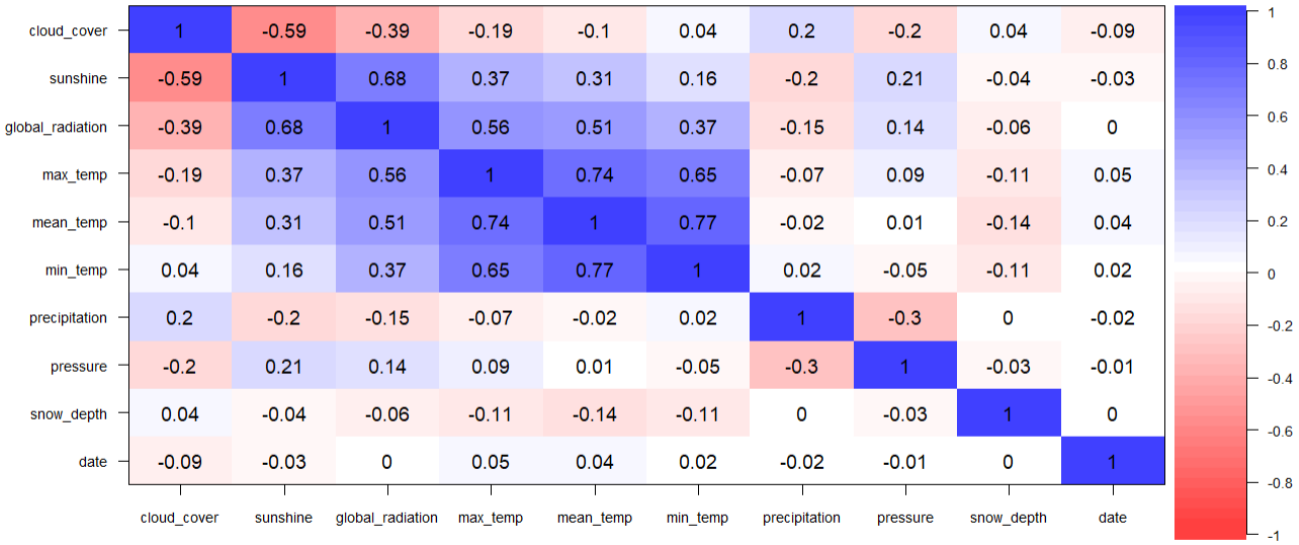
Macierz korelacji zmiennych po wygenerowani braków za pomocą mechanizmu MCAR:

		A	B	C	D	E	F	G	H	I
	r(i)	**cloud_cover**	**sunshine**	**global_radiation**	**max_temp**	**mean_temp**	**min_temp**	**precipitation**	**pressure**	**snow_depth**
A	**cloud_cover**	1	-0,7443	-0,4917	-0,2324	-0,1242	0,0511	0,2397	-0,2471	0,0436
B	**sunshine**	-0,7443	1	0,8395	0,4703	0,3886	0,1955	-0,2466	0,2613	-0,0477
C	**global_radiation**	-0,4917	0,8395	1	0,6979	0,6318	0,4670	-0,1821	0,1800	-0,0756
D	**max_temp**	-0,2324	0,4703	0,6979	1	0,9237	0,8182	-0,0943	0,1077	-0,1426
E	**mean_temp**	-0,1242	0,3886	0,6318	0,9237	1	0,9546	-0,0268	0,0161	-0,1644
F	**min_temp**	0,0511	0,1955	0,4670	0,8182	0,9546	1	0,0277	-0,0566	-0,1658
G	**precipitation**	0,2397	-0,2466	-0,1821	-0,0943	-0,0268	0,0277	1	-0,3725	-0,0052
H	**pressure**	-0,2471	0,2613	0,1800	0,1077	0,0161	-0,0566	-0,3725	1	-0,0390
I	**snow_depth**	0,0436	-0,0477	-0,0756	-0,1426	-0,1644	-0,1658	-0,0052	-0,0390	1

Po wygenerowaniu braków danych wspomniane wysokie korelacje pomiędzy zmiennymi wciąż są obecne. Korelacja pomiędzy zmiennymi dotyczącymi pomiarów opadu deszczu i śniegu wzrosła, jednak wciąż jest bardzo niska i ma znak ujemny.

Macierz korelacji po imputacji metodą k najbliższych sąsiadów (z odległością Gowera):

Macierz korelacji po imputacji regresją:



Macierz korelacji po imputacji metodą Random forest:

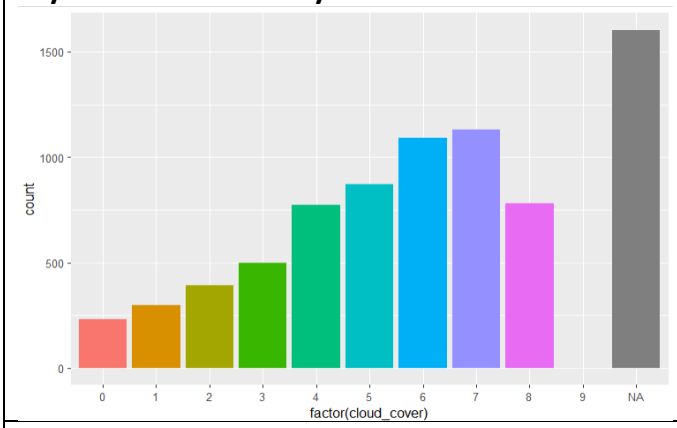
Na podstawie analizy korelogramów można zauważyć, że dla zmiennych cloud_cover, sunshine, global_radiation, min_temp, preassure najlepsza okazała się imputacja metodą k najbliższych sąsiadów – wartości te są najbliższe pełnemu zbiorowi danych. Dla zmiennej precipitation bardzo podobne wyniki daje zarówno imputacja metodą k najbliższych sąsiadów oraz lasów losowych. W przypadku zmiennych max_temp, mean_temp, snow_depth najlepsze wyniki daje metoda lasów losowych. Ponownie najgorzej wypadła metoda imputacji regresją, gdzie wyniki najbardziej odbiegają od tych dla całego zbioru danych.

Wizualizacja wybranych zmiennych przed imputacją braków wywołanych mechanizmem MCAR i po imputacji

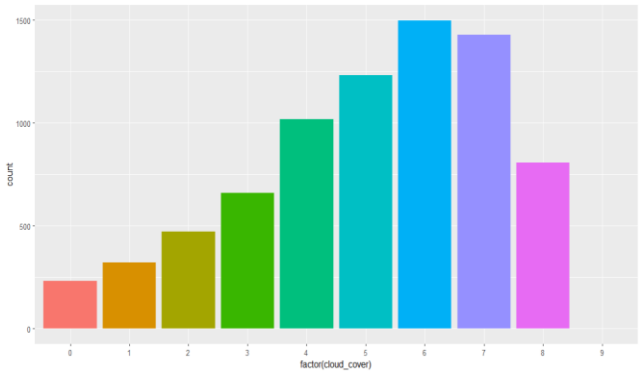
We wszystkich wizualizacjach pełnego zbioru danych, zbioru danych z brakami oraz zbiorów danych po imputacjach będą porównywane te same zmienne w celu zaobserwowania różnic

Zmienna cloud_cover:

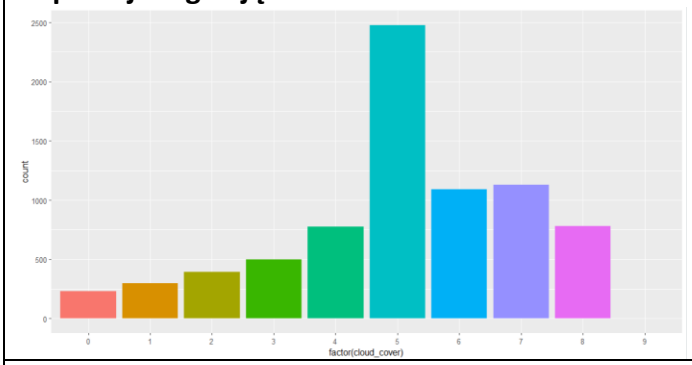
Wykres z brakami danych:



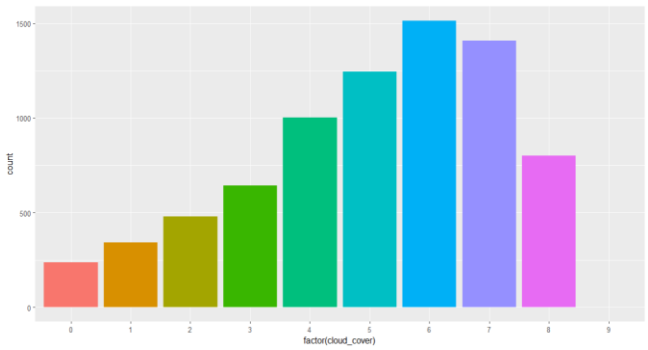
Imputacja metodą k najbliższych sąsiadów (z odległością Gowera):



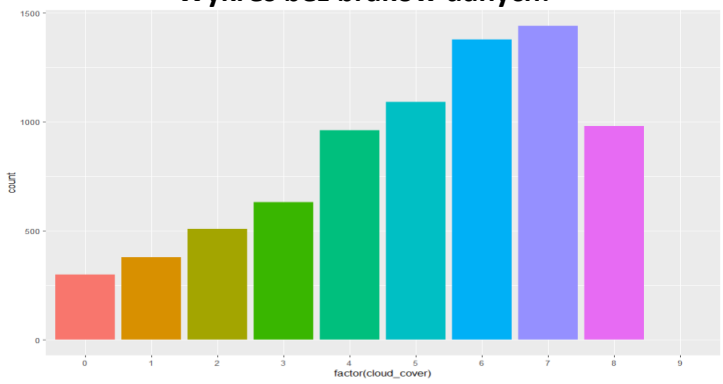
Imputacja regresją:



Imputacja metodą Random Forest:



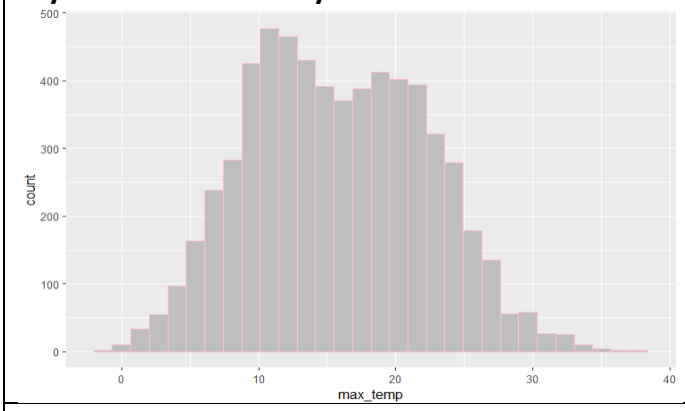
Wykres bez braków danych:



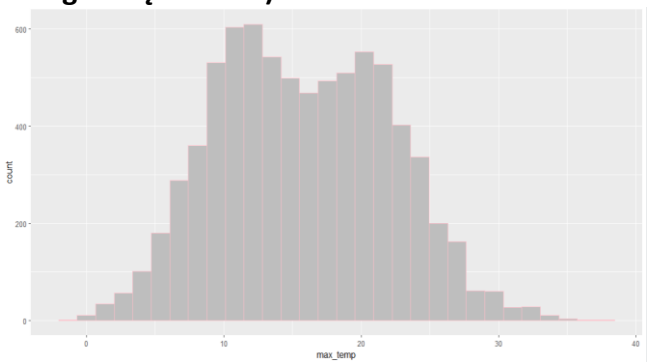
Analizując wykresy zmiennej cloud_cover przed i po imputacji braków danych można zauważyć, że tutaj również najbardziej obiega ten, gdzie braki danych zostały zimputowane za pomocą regresji. W przypadku parametrów 0-4 dla zmiennej, wyniki są podobne na trzech wykresach (imputacja k najbliższych sąsiadów, lasów losowych i wykres bez braków). Dla wartości zmiennej równej 5 i 6 zostało przydzielono nieco więcej wartości, w porównaniu do pełnego zbioru danych, natomiast dla wartości 7 i 8 mniej. Wartość 9 stale pozostała w bardzo małej ilości. Wykres gdzie braki zostały uzupełnione metodą k najbliższych sąsiadów najbardziej odwzorowuje wykres wykonany dla pełnego zbioru danych.

Zmienna max_temp:

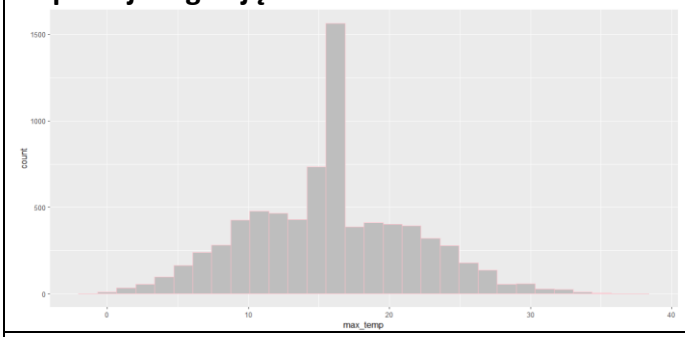
Wykres z brakami danych:



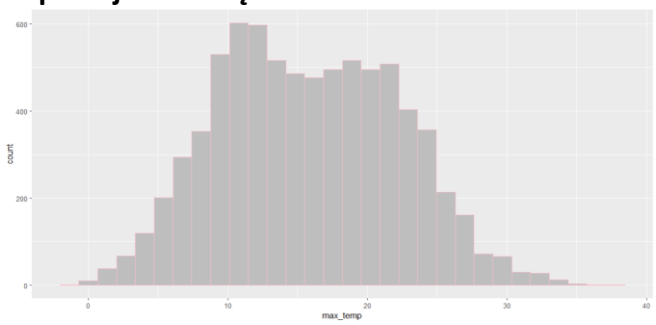
Imputacja metodą k najbliższych sąsiadów (z odległością Gowera):



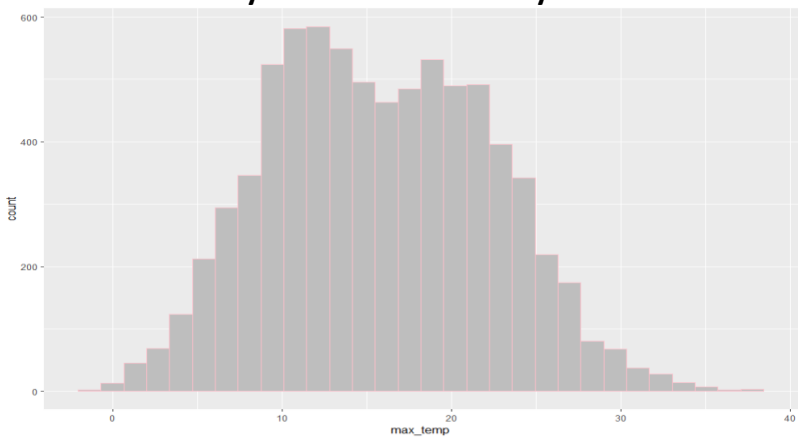
Imputacja regresją:



Imputacja metodą Random Forest:



Wykres bez braków danych:



Analizując wykresy zmiennej max_temp przed i po imputacji braków danych można zauważyć, że najbardziej obiega ten, gdzie braki danych zostały zimputowane za pomocą regresji – w żadnym stopniu nie przypomina wykresu dla pełnego zbioru danych. W tym przypadku również bardzo zbliżone wyniki dają metody k najbliższych sąsiadów i lasów losowych – oba wykresy są bliskie wykresowi pełnego zbioru danych.

Wnioski:

Zarówno metoda imputacji k najbliższych sąsiadów jak i lasów losowych okazały się odpowiednie dla tego typu zmiennych. Ponieważ metoda regresji wypadła najgorzej – wyniki zdecydowanie odbiegają od wartości pełnego zbioru danych, przy kolejnym mechanizmie losowania braków zostanie w zamian zastosowana imputacja metodą Hot-deck.

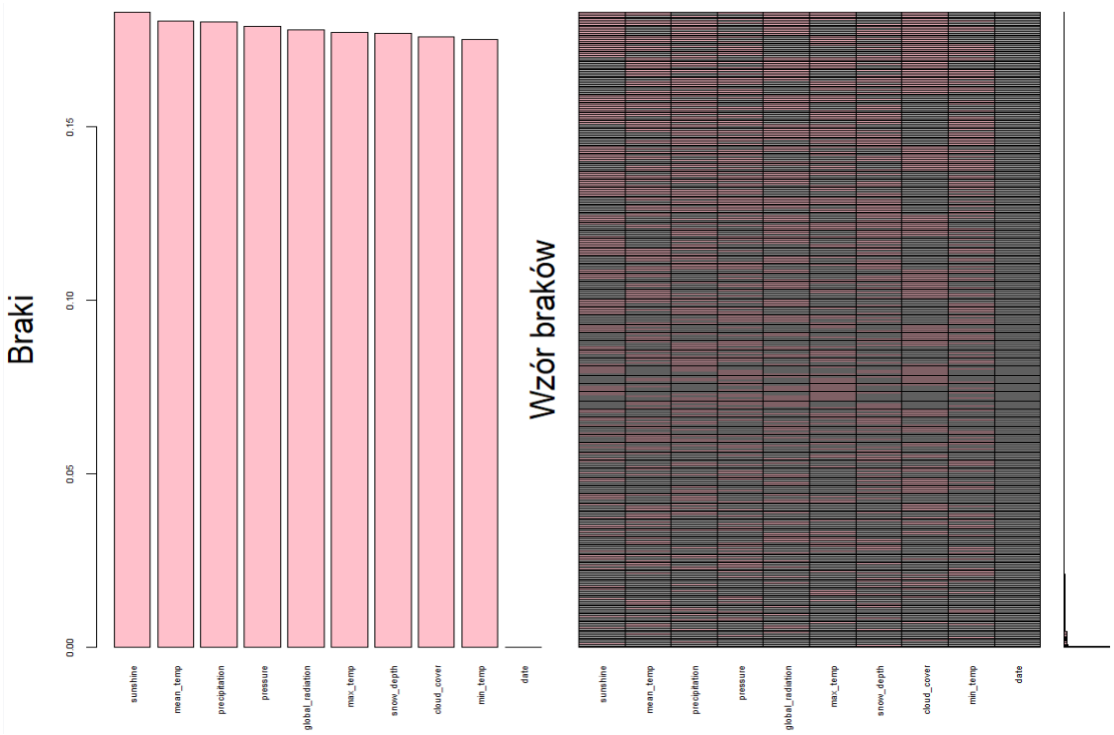
2. MAR precipitation 0: 25%; 1:10%

Mechanizm MAR (ang. *Missing at Random*) polega na losowym występowaniu braków. Zachodzi wówczas, gdy prawdopodobieństwo wystąpienia braku danych w obrębie zmiennej Y jest niezależne od wartości tej zmiennej Y, ale jest zależne od wartości innej zmiennej X. W takim przypadku zakładamy, że nie ma powiązania występowania braków z wartościami danej zmiennej.

Prezentacja wygenerowanych braków:

	A	B	C	D	E	F	G	H	I	J
1	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
2	braki	0	18	11	7	5	0	102450	0	20000101
3	7	1	20	12	8	5	0	102530	braki	20000102
4	8	0	13	10	9	7	6	101860	0	20000103
5	5	3	braki	11	7	braki	braki	101480	braki	20000104
6	braki	1	25	11	6	2	1	101420	0	20000105
7	braki	1	20	braki	9	7	braki	101270	0	20000106
8	6	2	31	9	7	3	2	101720	0	20000107
9	braki	6	braki	braki	7	6	0	101650	0	20000108
10	0	braki	55	8	braki	-1	braki	102760	0	20000109
11	5	4	40	10	2	-3	0	103470	0	20000110
12	7	0	16	8	5	braki	0	braki	0	20000111
13	7	0	14	5	8	8	3	102340	0	20000112
14	7	0	14	6	4	braki	0	101540	0	20000113
15	braki	4	45	7	3	braki	0	102330	0	20000114
16	6	1	25	7	5	3	0	braki	0	20000115
17	6	4	44	9	5	braki	0	braki	braki	20000116
18	5	braki	44	8	4	braki	0	103790	0	20000117
19	braki	0	15	7	braki	1	0	103540	braki	20000118
20	6	1	24	5	5	braki	braki	103700	0	20000119

Wizualizacja braków MAR:



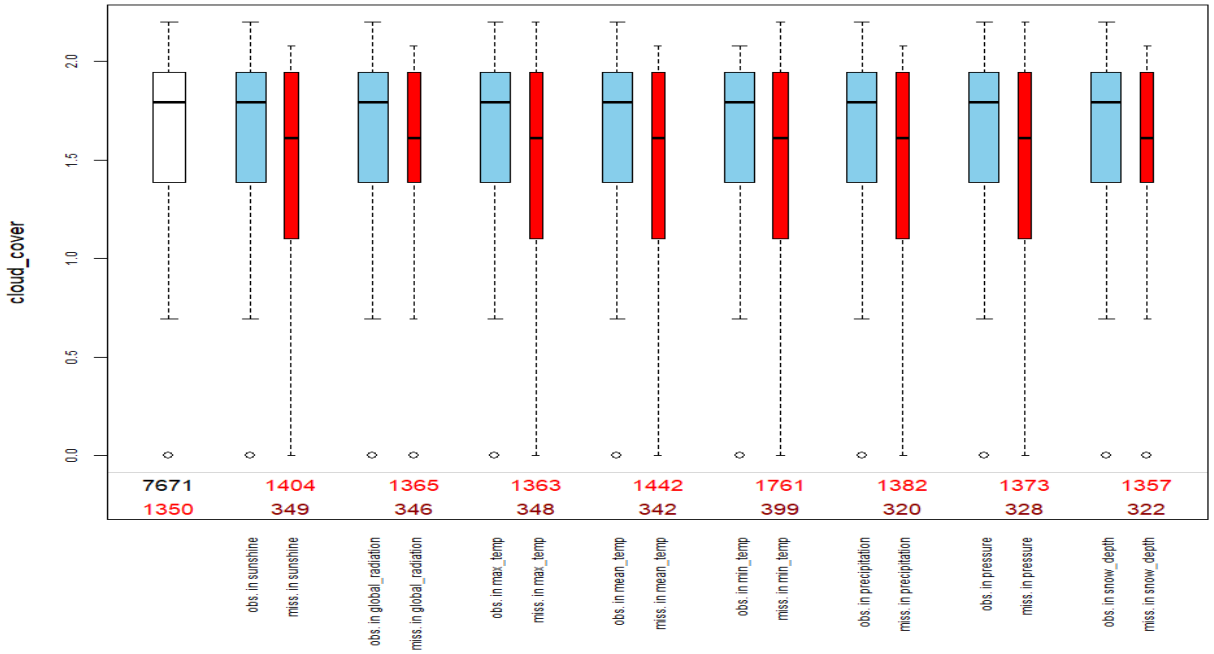
Ilości braków w zbiorze dla poszczególnych zmiennych:

```
  cloud_cover      sunshine global_radiation      max_temp     mean_temp      min_temp
         1350          1404             1365          1359          1385          1343
precipitation      pressure    snow_depth          date
         1382          1373          1357             0
```

Zmienne uszeregowane w kolejności od największego do najmniejszego udziału braków w zbiorze danych:

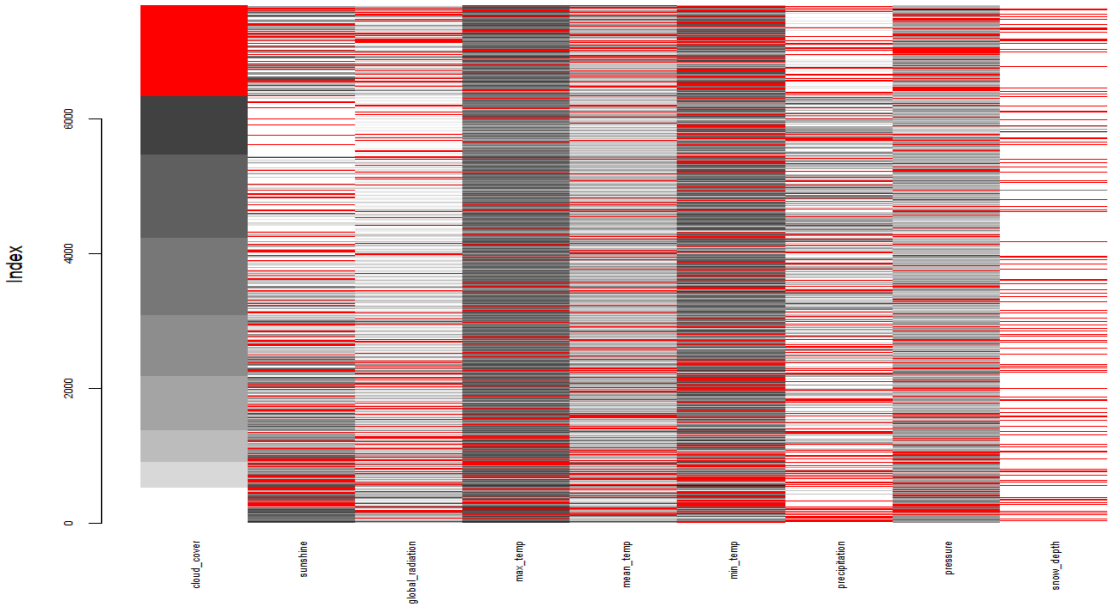
Variable	Count
sunshine	0.1830270
mean_temp	0.1805501
precipitation	0.1801590
pressure	0.1789858
global_radiation	0.1779429
max_temp	0.1771607
snow_depth	0.1769000
cloud_cover	0.1759875
min_temp	0.1750750
date	0.0000000

Wykresy pudełkowe braków – zlogarytmowane wartości obserwowane i braki wartości dla poszczególnych zmiennych:



Ilość braków nie są równomiernie rozłożone, choć dla niektórych zmiennych występuje bardzo zbliżona ilość braków.

Rozkład braków i wartości:



Prezentacja imputacji poszczególnymi metodami:

Zmienne cloud_cover i sunshine zostały zaokrąglone do liczb całkowitych, ponieważ tak były wyrażane w zbiorze danych pierwotnych

1) k najbliższych sąsiadów (z odległością Gowera)

MARkNN<-kNN(MAR, numFun = weightedMean, weightDist=TRUE)

MARkNN<-MARkNN[,-c(11:20)]

	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
1	8	0	18.00000	10.800000	7.000000	4.9000000	0.00000000	102450.0	0.0000000	20000101
2	7	1	20.00000	11.500000	7.900000	5.0000000	0.20000000	102530.0	0.0000000	20000102
3	8	0	13.00000	9.500000	9.400000	7.2000000	6.00000000	101860.0	0.0000000	20000103
4	5	3	121.69594	11.000000	7.000000	6.3753722	0.44138196	101480.0	0.0000000	20000104
5	7	1	25.00000	10.800000	6.400000	1.9000000	0.80000000	101420.0	0.0000000	20000105
6	7	1	20.00000	12.486009	8.900000	7.0000000	7.03132115	101270.0	0.0000000	20000106
7	6	2	31.00000	9.200000	7.200000	3.4000000	2.00000000	101720.0	0.0000000	20000107
8	5	6	148.96495	11.790485	7.400000	5.7000000	0.00000000	101650.0	0.0000000	20000108
9	0	5	55.00000	7.800000	3.559057	-0.7000000	0.83500049	102760.0	0.0000000	20000109
10	5	4	40.00000	10.200000	2.200000	-3.3000000	0.00000000	103470.0	0.0000000	20000110

2) Hot-deck

MAR_HD <- hotdeck(MAR,ord_var="date")

MAR_HD<-MAR_HD [,-c(11:19)]

	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
1	7	0	18	10.8	7.0	4.9	0.0	102450	0	20000101
2	7	1	20	11.5	7.9	5.0	0.2	102530	0	20000102
3	8	0	13	9.5	9.4	7.2	6.0	101860	0	20000103
4	5	3	13	11.0	7.0	7.2	6.0	101480	0	20000104
5	5	1	25	10.8	6.4	1.9	0.8	101420	0	20000105
6	5	1	20	10.8	8.9	7.0	0.8	101270	0	20000106
7	6	2	31	9.2	7.2	3.4	2.0	101720	0	20000107
8	6	6	31	9.2	7.4	5.7	0.0	101650	0	20000108
9	0	6	55	7.8	7.4	-0.7	0.0	102760	0	20000109
10	5	4	40	10.2	2.2	-3.3	0.0	103470	0	20000110

3) Random Forest

MAR_RandF<-rangerImpute(cloud_cover+sunshine+global_radiation+min_temp+mean_temp+

max_temp+precipitation+pressure+snow_depth~date,data=MAR)

MAR_RandF<-MAR_RandF[,-c(11:19)]

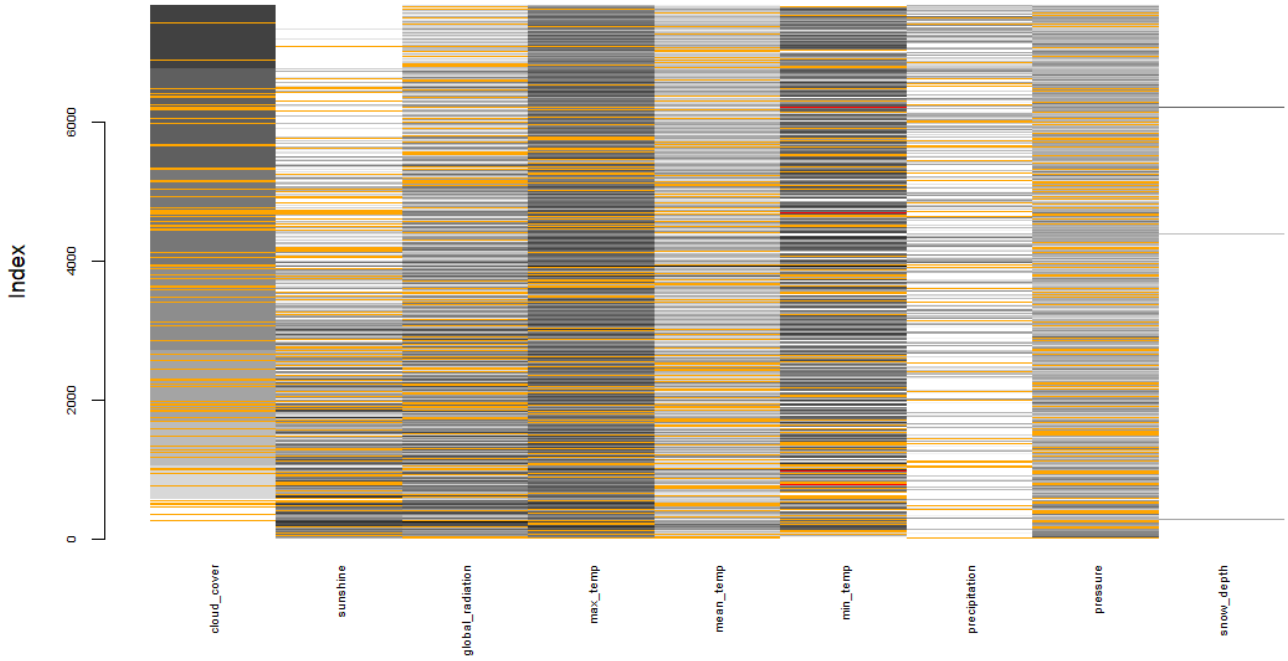
	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
1	7	0	18.00000	10.800000	7.000000	4.9000000	0.0000000	102450.00	0	20000101
2	7	1	20.00000	11.500000	7.900000	5.0000000	0.2000000	102530.00	0	20000102
3	8	0	13.00000	9.500000	9.400000	7.2000000	6.0000000	101860.00	0	20000103
4	5	3	19.55087	11.000000	7.000000	5.4256767	2.8770133	101480.00	0	20000104
5	5	1	25.00000	10.800000	6.400000	1.9000000	0.8000000	101420.00	0	20000105
6	5	1	20.00000	10.288157	8.900000	7.0000000	1.3845767	101270.00	0	20000106
7	6	2	31.00000	9.200000	7.200000	3.4000000	2.0000000	101720.00	0	20000107
8	5	6	35.50523	9.341023	7.400000	5.7000000	0.0000000	101650.00	0	20000108
9	0	4	55.00000	7.800000	6.347980	-0.7000000	0.4361767	102760.00	0	20000109
10	5	4	40.00000	10.200000	2.200000	-3.3000000	0.0000000	103470.00	0	20000110

Zestawienie statystyk opisowych zbiorów danych po wykonanych imputacjach do pełnego zbioru danych oraz zbioru danych z wygenerowanymi brakami:

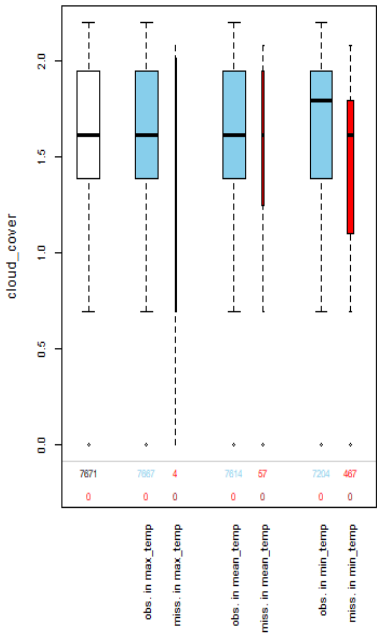
	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth
	Pełen zbiór danych								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	5,06	4,33	118,79	15,81	11,91	8	1,73	101516,7	0,02
Me	5	3,5	95	15,5	11,8	8,2	0	101610	0
s	2,21	3,99	88,53	6,52	5,67	5,24	3,69	1046,43	0,28
Vs	0,44	0,92	0,75	0,41	0,48	0,66	2,13	0,01	13,81
g1	-0,58	0,71	0,67	0,14	0	-0,17	3,79	-0,42	22,12
s(x_śr)	0,98	1,92	8,12	1,64	1,64	1,85	2,81	3,28	1,95
	Zbiór danych z brakami wywołanymi mechanizmem MAR								
n	6321	6267	6306	6312	6286	6328	6289	6298	6314
x_śr	5,15	4,18	116,61	15,69	11,88	8,06	2,08	101459,7	0,02
Me	6	3,3	92	15,4	11,8	8,2	0,2	101550	0
s	2,19	3,97	87,87	6,43	5,63	5,22	3,99	1057,8	0,29
Vs	0,43	0,95	0,75	0,41	0,47	0,65	1,92	0,01	14,35
g1	-0,64	0,76	0,7	0,15	0	-0,17	3,41	-0,4	22,42
s(x_śr)	0,97	1,94	8,14	1,62	1,63	1,84	2,77	3,32	2,04
	Imputacja metodą k najbliższych sąsiadów (z odległością Gowera)								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	5,05	4,32	119,02	15,81	11,92	8	1,69	101507,6	0,02
Me	5	4	94	15,5	11,8	8,3	0	101590	0
s	2,21	4,03	88,8	6,55	5,66	5,24	3,59	1054,31	0,28
Vs	0,44	0,93	0,75	0,41	0,47	0,65	2,12	0,01	14,2
g1	-0,57	0,7	0,67	0,14	0	-0,17	3,8	-0,43	22,25
s(x_śr)	0,98	1,94	8,14	1,65	1,64	1,85	2,76	3,31	1,98
	Imputacja metodą Hot-deck								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	5,14	4,22	118,34	15,80	11,92	8,01	1,99	101504,04	0,02
Me	6	3	94	15,5	11,8	8,1	0,2	101600	0
s	2,19	4,01	88,52	6,49	5,65	5,23	3,91	1052,53	0,3
Vs	0,43	0,95	0,75	0,41	0,47	0,65	1,97	0,01	12,59
g1	-0,63	0,73	0,67	0,13	-0,01	-0,16	3,6	-0,42	19,59
s(x_śr)	0,97	1,95	8,14	1,63	1,64	1,85	2,77	3,3	1,93
	Imputacja metodą Random Forest								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	5,12	4,22	118,17	15,79	11,93	8,02	1,99	101501,62	0,02
Me	5	3	97	15,5	11,8	8,2	0,21	101600	0
s	2,1	3,84	86,89	6,42	5,62	5,19	3,75	1023,54	0,28
Vs	0,41	0,91	0,74	0,41	0,47	0,65	1,88	0,01	13,07
g1	-0,62	0,73	0,65	0,12	-0,01	-0,15	3,55	-0,44	21,29
s(x_śr)	0,93	1,87	7,99	1,62	1,63	1,83	2,66	3,21	1,92

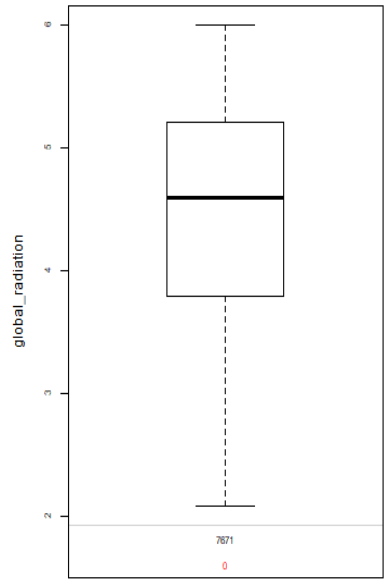
Na podstawie powyższej tabeli można zauważyć, że wszystkie metody imputacji działają poprawnie, ponieważ nie występują żadne braki. Najwięcej brakujących danych zostało wylosowanych dla zmiennych sunshine i mean_temp. Kolorem zielonym zostały zaznaczone wartości statystyk opisowych, które są takie same dla pełnego zbioru danych jak i po konkretnej imputacji. W przypadku losowania braków mechanizmem MAR, również najlepsza okazała się imputacja metodą k najbliższych sąsiadów, kolejno lasów losowych i na trzecim miejscu Hot-deck, jednak wszystkie dają stosunkowo dobre wyniki. Żadne z parametrów nie odbiegają daleko od statystyk opisowych pełnego zbioru danych.

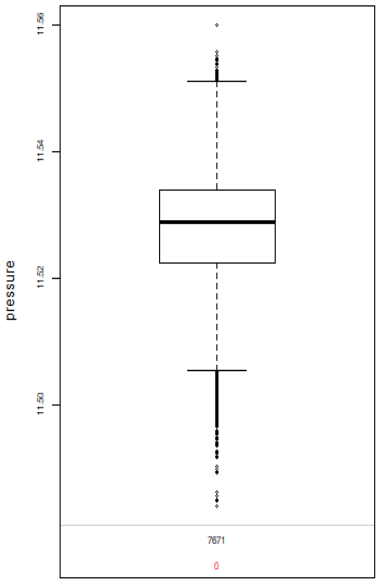
Wizualizacja rozkładu braków po imputacji metodą k najbliższych sąsiadów (z odległością Gowera):

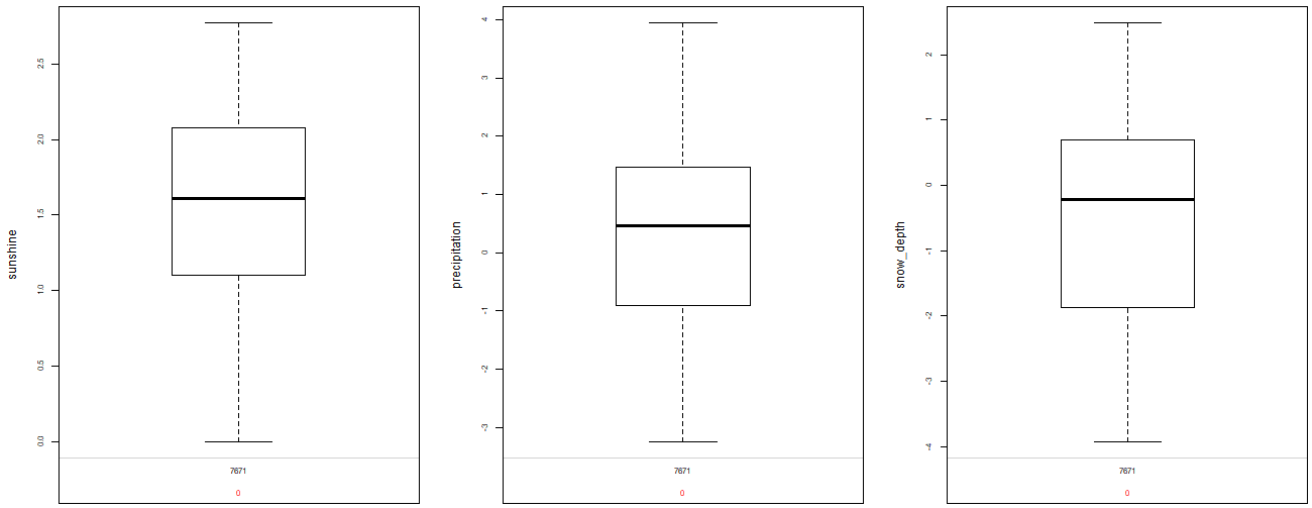


Wykresy pudełkowe po imputacji metodą k najbliższych sąsiadów (z odległością Gowera):







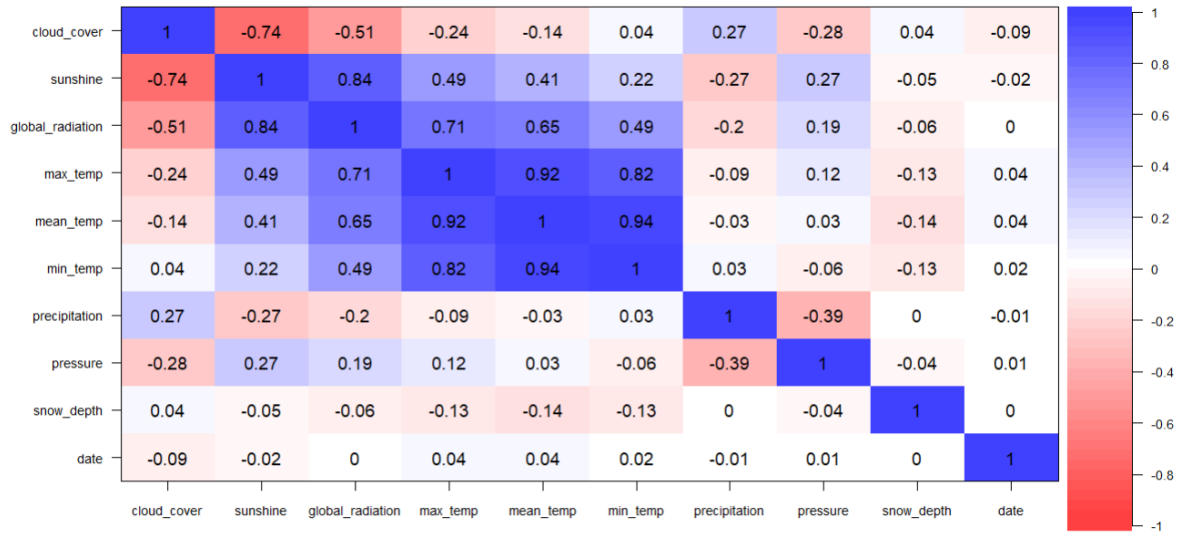


Macierz korelacji zmiennych po wygenerowani braków za pomocą mechanizmu MAR:

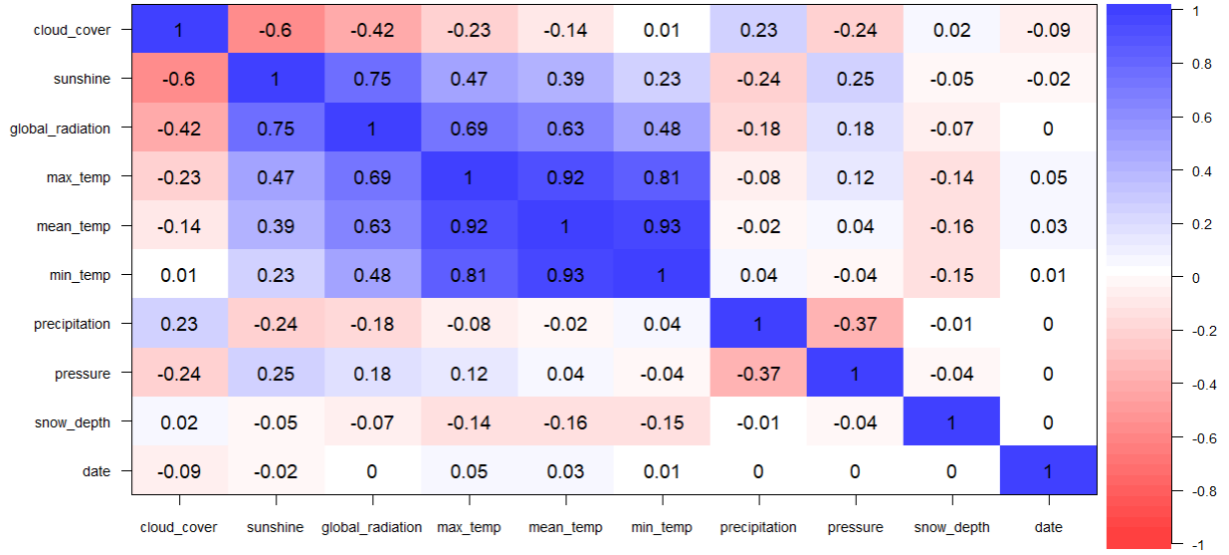
		A	B	C	D	E	F	G	H	I
	r(i)	**cloud_cover**	**sunshine**	**global_radiation**	**max_temp**	**mean_temp**	**min_temp**	**precipitation**	**pressure**	**snow_depth**
A	**cloud_cover**	1	-0,7457	-0,4908	-0,2278	-0,1265	0,0489	0,2691	-0,2633	0,0290
B	**sunshine**	-0,7457	1	0,8472	0,4667	0,3918	0,1952	-0,2709	0,2761	-0,0417
C	**global_radiation**	-0,4908	0,8472	1	0,6883	0,6289	0,4586	-0,1983	0,1897	-0,0583
D	**max_temp**	-0,2278	0,4667	0,6883	1	0,9195	0,8103	-0,0863	0,1297	-0,1263
E	**mean_temp**	-0,1265	0,3918	0,6289	0,9195	1	0,9538	-0,0287	0,0432	-0,1463
F	**min_temp**	0,0489	0,1952	0,4586	0,8103	0,9538	1	0,0321	-0,0394	-0,1368
G	**precipitation**	0,2691	-0,2709	-0,1983	-0,0863	-0,0287	0,0321	1	-0,3848	-0,0055
H	**pressure**	-0,2633	0,2761	0,1897	0,1297	0,0432	-0,0394	-0,3848	1	-0,0372
I	**snow_depth**	0,0290	-0,0417	-0,0583	-0,1263	-0,1463	-0,1368	-0,0055	-0,0372	1

Po wygenerowaniu braków wspomniane wysokie korelacje pomiędzy zmiennymi wciąż są obecne. Korelacja pomiędzy zmiennymi dotyczącymi pomiarów opadu deszczu i śniegu wzrosła jednak wciąż jest bardzo niska i ma ujemny znak.

Macierz korelacji po imputacji metodą k najbliższych sąsiadów (z odległością Gowera):



Macierz korelacji po imputacji metodą hot-deck:



Macierz korelacji po imputacji metodą Random forest:

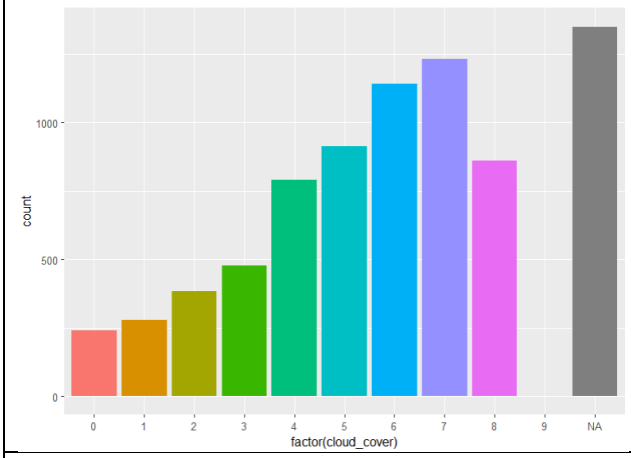
Na podstawie analizy korelogramów można zauważyć, że w przypadku wysokich korelacji bardziej zbliżone do tych pełnego zbioru danych są te, gdzie braki danych zostały zimputowane metodą k najbliższych sąsiadów, natomiast dla niskich wartości korelacji bardziej zbliżone do tych, dla pełnego zbioru danych są te, gdzie braki danych zostały zimputowane metodą lasów losowych. Jednak podsumowując najbardziej zbliżone wartości korelacji do tych z pełnego zbioru danych przypadają dla danych zimputowanych metodą k najbliższych sąsiadów. Korelacje w zbiorze danych po imputacji braków metodą Hot-deck wypadają najgorzej, jednak nie odstają znacząco od korelacji w zbiorze danych po imputacji braków metodą lasów losowych.

Wizualizacja wybranych zmiennych przed imputacją braków wywołanych mechanizmem MAR i po imputacji

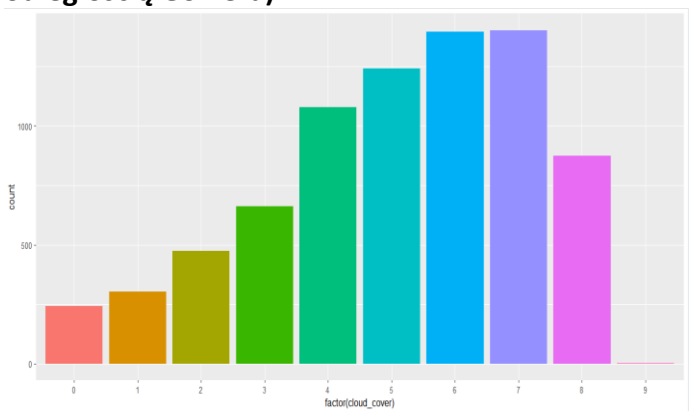
We wszystkich wizualizacjach pełnego zbioru danych, zbioru danych z brakami oraz zbiorów danych po imputacjach będą porównywane te same zmienne w celu zaobserwowania różnic

Zmienna cloud_cover:

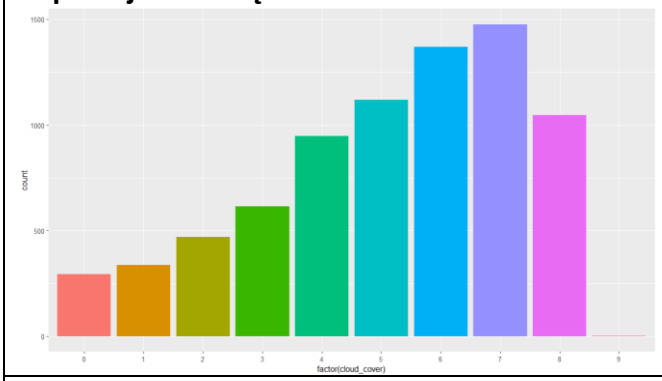
Wykres z brakami danych:



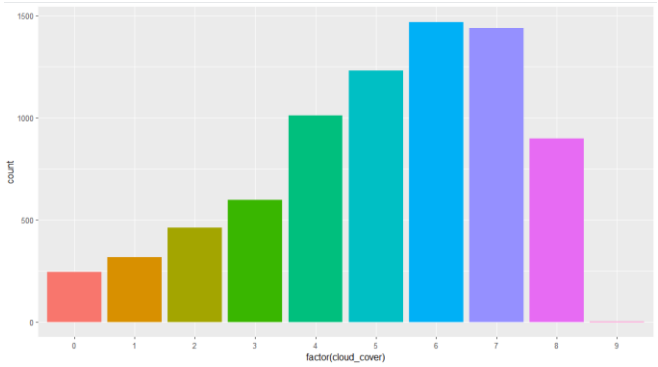
Imputacja metodą k najbliższych sąsiadów (z odległością Gowera):



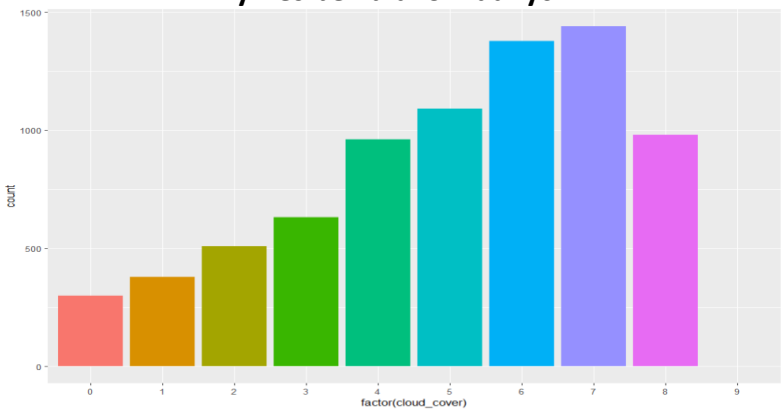
Imputacja metodą Hot-deck:



Imputacja metodą Random Forest:



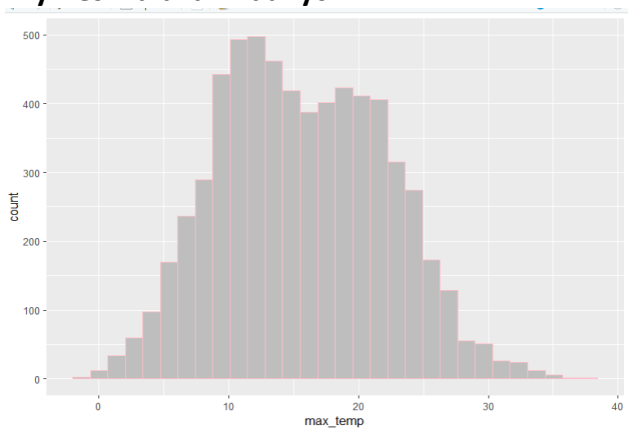
Wykres bez braków danych:



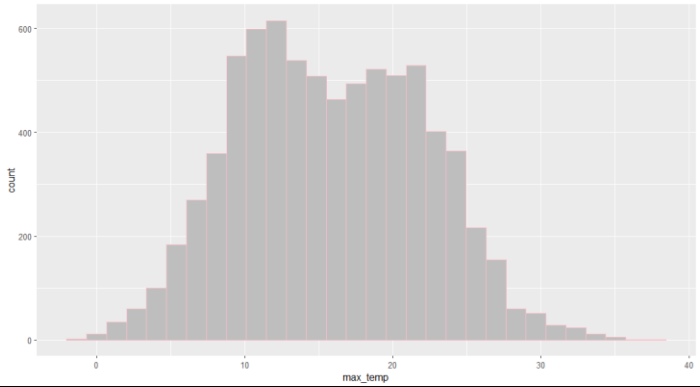
Analizując wykresy zmiennej cloud_cover przed i po imputacji braków danych można zauważyć, że patrząc na samą wizualizację zmiennych, najbardziej zbliżony wykres do tego wykonanego na pełnym zbiorze danych, to ten wykonany na danych, gdzie braki zimputowano metodą Hot-deck. Kolejno najbardziej zbliżony do wykresu bez braków danych jest wykres danych po imputacji braków metodą lasów losowych i na końcu metodą k najbliższych sąsiadów.

Zmienna max_temp:

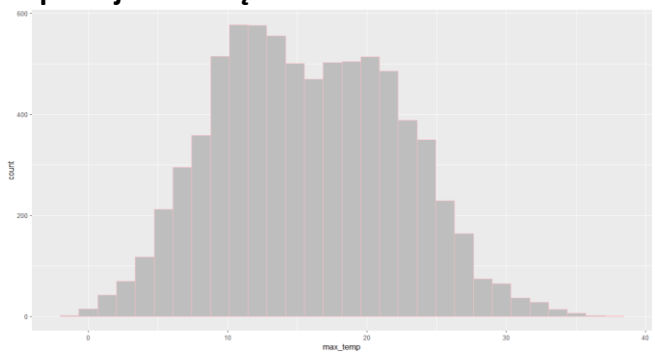
Wykres z brakami danych:



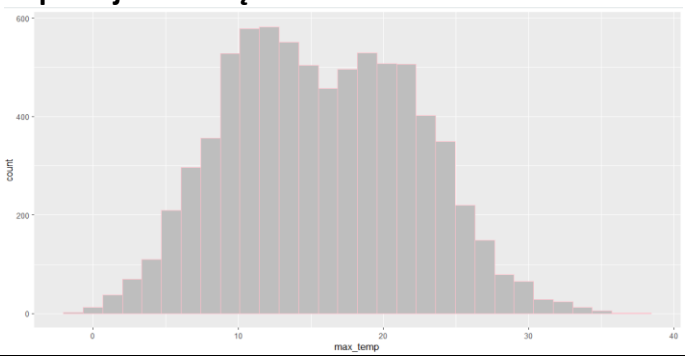
Imputacja metodą k najbliższych sąsiadów (z odległością Gowera):



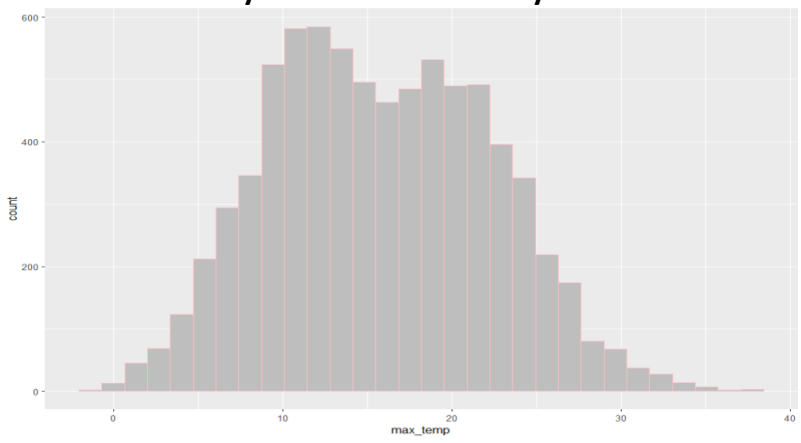
Imputacja metodą Hot-deck:



Imputacja metodą Random Forest:



Wykres bez braków danych :



Analizując wykresy zmiennej max_temp przed i po imputacji braków danych można zauważyć, że najbardziej obiega od pozostałych ten wykonany na podstawie danych, gdzie braki zimputowano za pomocą metody Hot-deck. Najbliższy wykresowi wykonanemu na pełnym zbiorze danych jest wykres po imputacji braków metodą lasów losowych. Podsumowując wszystkie wykresy dobrze odwzorowują wykres zawierający pełne dane.

Wnioski:

Przy mechanizmie losowania braków MAR wszystkie metody okazały się odpowiednie, jednak najlepszą jest metoda k najbliższych sąsiadów, a kolejno lasów losowych. Ponieważ wszystkie metody okazały się trafne w kolejnym mechanizmie – NMAR, również zostaną one zastosowane.

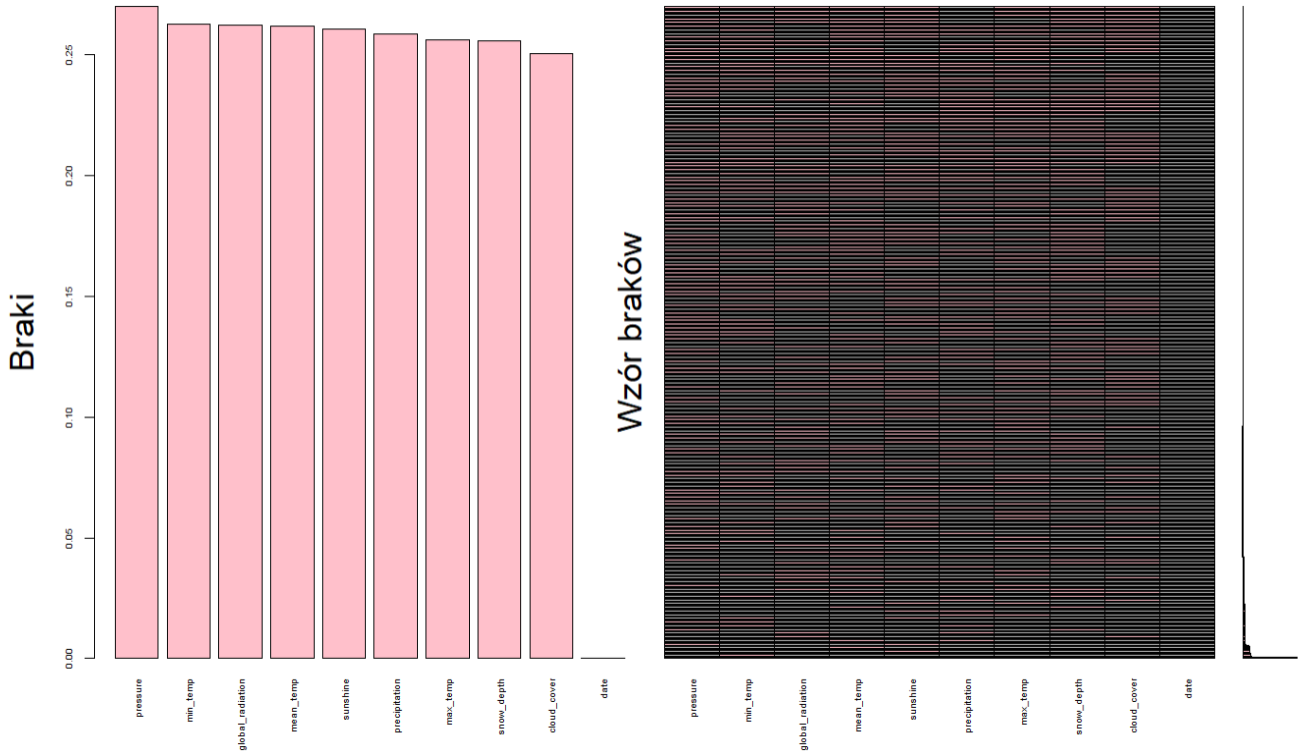
3. Mechanizm NMAR PPX cloud_cover 20%

Są to braki nie losowe (ang. *Missing Not at Random*). Gdy dane są NMAR, fakt braku danych jest systematycznie powiązany z danymi nieobserwowanymi, czyli brak jest związany ze zdarzeniami lub czynnikami, które nie są mierzone przez badacza.

Prezentacja wygenerowanych braków:

	A	B	C	D	E	F	G	H	I	J
1	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
2	7	braki	braki	braki	braki	5	0	102450	0	20000101
3	7	1	braki	braki	8	braki	0	braki	0	20000102
4	8	braki	braki	10	9	braki	braki	101860	0	20000103
5	braki	3	34	11	7	4	0	braki	0	20000104
6	5	1	25	braki	6	2	1	101420	0	20000105
7	6	braki	20	braki	braki	7	braki	braki	0	20000106
8	braki	braki	31	9	7	3	2	101720	0	20000107
9	braki	6	52	7	7	6	0	101650	0	20000108
10	0	7	55	8	3	-1	0	102760	0	20000109
11	5	4	40	10	2	-3	0	103470	0	20000110
12	braki	0	braki	braki	5	1	0	103240	braki	20000111
13	7	0	14	braki	braki	8	3	braki	braki	20000112
14	7	0	14	6	4	3	0	braki	braki	20000113
15	6	4	45	7	braki	braki	0	102330	braki	20000114
16	6	1	25	braki	5	3	0	103640	braki	20000115
17	6	braki	44	9	5	braki	0	104050	0	20000116
18	5	braki	44	braki	4	braki	0	103790	braki	20000117
19	6	0	15	7	5	braki	0	103540	braki	20000118
20	braki	1	braki	5	5	4	0	103700	braki	20000119

Wizualizacja braków NMAR:



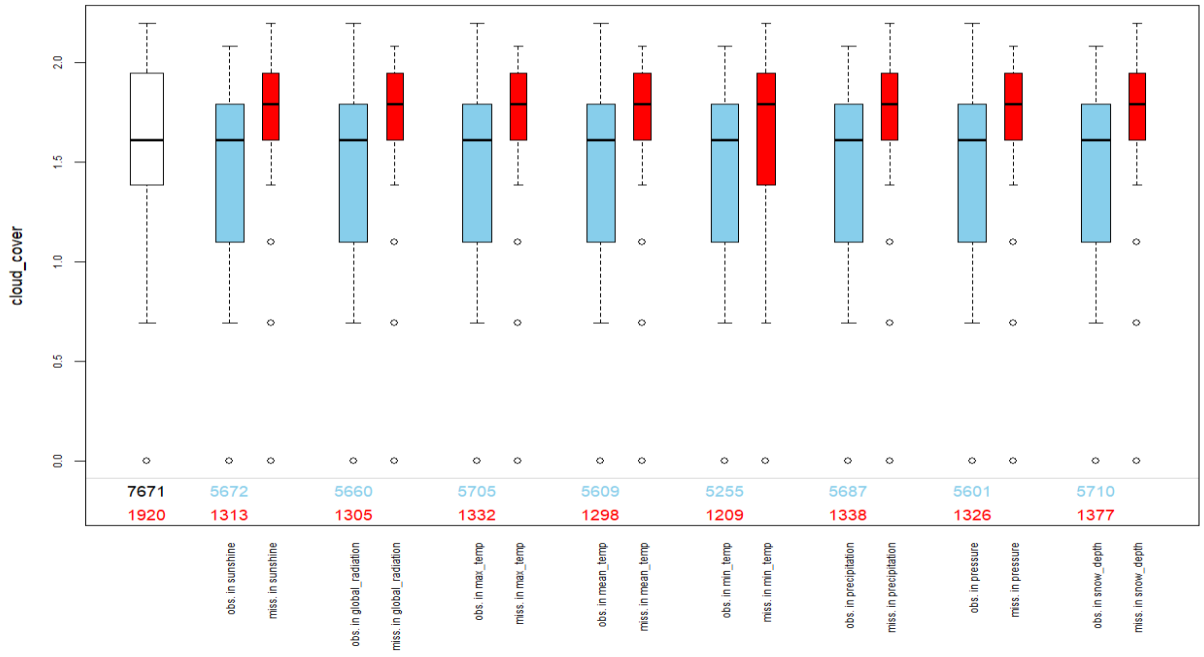
Ilości braków w zbiorze dla poszczególnych zmiennych:

```
cloud_cover        sunshine global_radiation        max_temp       mean_temp
       1920            1999             2011            1963            2007
   min_temp   precipitation        pressure      snow_depth            date
       2015            1984             2070            1961               0
```

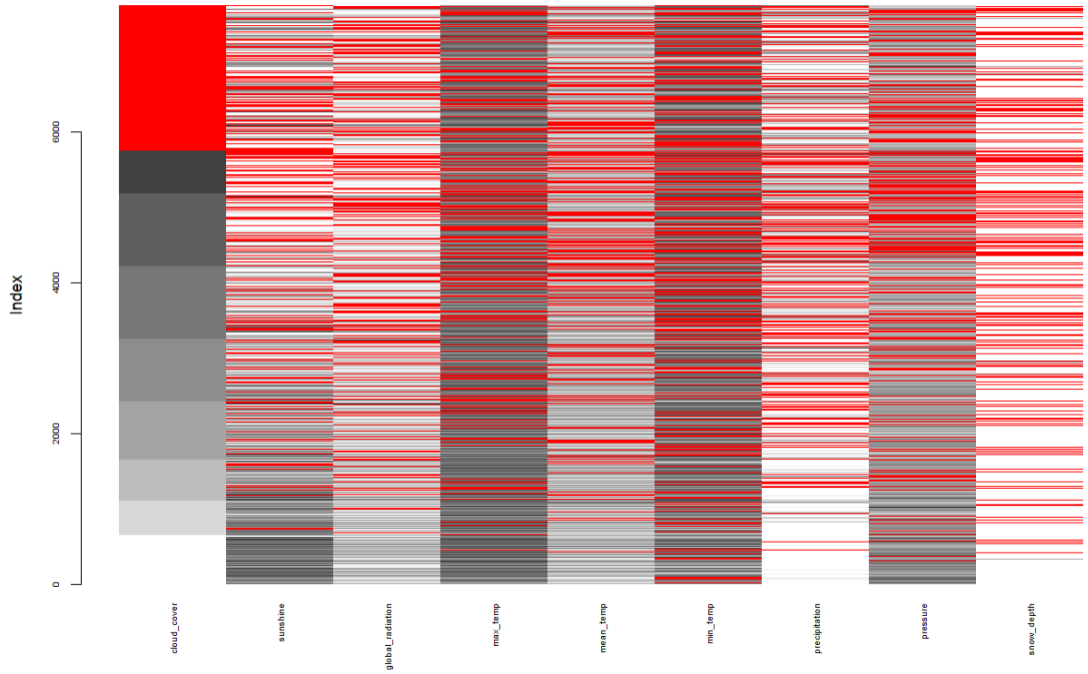
Zmienne uszeregowane w kolejności od największego do najmniejszego udziału braków w zbiorze danych:

Variable	Count
pressure	0.2698475
min_temp	0.2626776
global_radiation	0.2621562
mean_temp	0.2616347
sunshine	0.2605918
precipitation	0.2586364
max_temp	0.2558988
snow_depth	0.2556381
cloud_cover	0.2502933
date	0.0000000

Wykresy pudełkowe braków – zlogarytmowane wartości obserwowane i braki wartości dla poszczególnych zmiennych:



Rozkład braków i wartości:



Prezentacja imputacji poszczególnymi metodami:

Zmienne cloud_cover i sunshine zostały zaokrąglone do liczb całkowitych, ponieważ tak były wyrażane w zbiorze danych pierwotnych

1) k najbliższych sąsiadów (z odległością Gowera)

NMARkNN<-kNN(NMAR, numFun = weightedMean, weightDist=TRUE)

NMARkNN<-NMARkNN[,-c(11:20)]

	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
1	7	3	91.27576	15.104544	11.391114	4.9000000	0.00000000	102450.00	0	20000101
2	7	1	60.64606	11.574904	7.900000	5.3132441	0.20000000	100843.19	0	20000102
3	8	2	74.62228	9.500000	9.400000	6.5700836	0.79981219	101860.00	0	20000103
4	5	3	34.00000	11.000000	7.000000	4.4000000	0.20000000	101764.67	0	20000104
5	5	1	25.00000	9.187995	6.400000	1.9000000	0.80000000	101420.00	0	20000105
6	6	2	20.00000	12.725573	8.571387	7.0000000	9.30089595	100365.98	0	20000106
7	4	6	31.00000	9.200000	7.200000	3.4000000	2.00000000	101720.00	0	20000107
8	3	6	52.00000	7.200000	7.400000	5.7000000	0.00000000	101650.00	0	20000108
9	0	7	55.00000	7.800000	3.200000	-0.7000000	0.20000000	102760.00	0	20000109
10	5	4	40.00000	10.200000	2.200000	-3.3000000	0.00000000	103470.00	0	20000110

2) Hot-deck

NMAR_HD <- hotdeck(NMAR,ord_var="date")

NMAR_HD<-NMAR_HD [,-c(11:19)]

	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
1	7	1	22	5.6	7.9	4.9	0.0	102450	0	20000101
2	7	1	34	5.6	7.9	4.9	0.2	102450	0	20000102
3	8	1	34	9.5	9.4	4.9	0.2	101860	0	20000103
4	8	3	34	11.0	7.0	4.4	0.2	101860	0	20000104
5	5	1	25	11.0	6.4	1.9	0.8	101420	0	20000105
6	6	1	20	11.0	6.4	7.0	0.8	101420	0	20000106
7	6	1	31	9.2	7.2	3.4	2.0	101720	0	20000107
8	6	6	52	7.2	7.4	5.7	0.0	101650	0	20000108
9	0	7	55	7.8	3.2	-0.7	0.2	102760	0	20000109
10	5	4	40	10.2	2.2	-3.3	0.0	103470	0	20000110

3) Random Forest

NMAR_RandF<-rangerImpute(cloud_cover+sunshine+global_radiation+min_temp+mean_temp+

max_temp+precipitation+pressure+snow_depth~date,data=NMAR)

NMAR_RandF<-NMAR_RandF[,-c(11:19)]

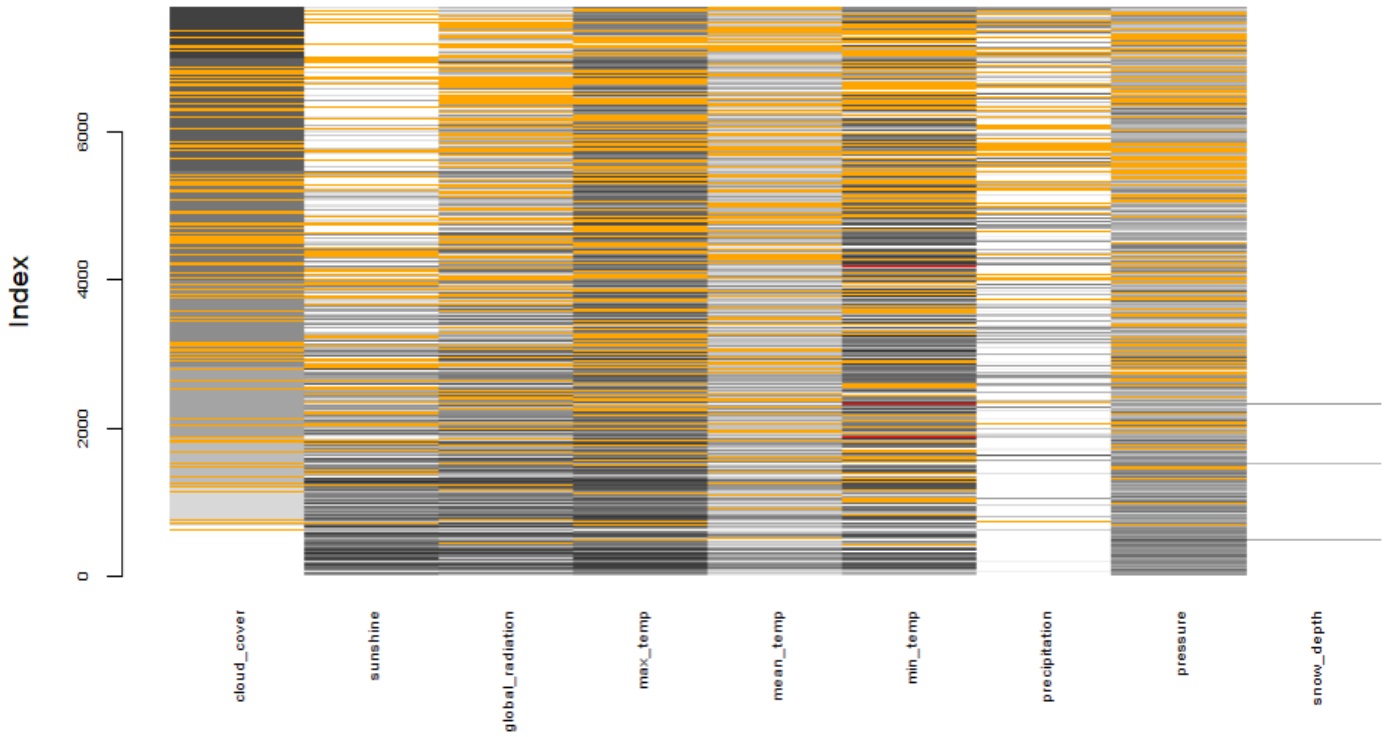
	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	date
1	7	3	30.41303	9.761550	8.185917	4.900000	0.000000000	102450.00	0	20000101
2	7	1	30.41303	9.761550	7.900000	4.215220	0.200000000	102030.04	0	20000102
3	8	3	30.41303	9.500000	9.400000	4.208220	0.469860000	101860.00	0	20000103
4	6	3	34.00000	11.000000	7.000000	4.400000	0.200000000	101816.84	0	20000104
5	5	1	25.00000	9.801150	6.400000	1.900000	0.800000000	101420.00	0	20000105
6	6	3	20.00000	9.445180	7.169910	7.000000	0.730860000	101758.75	0	20000106
7	4	4	31.00000	9.200000	7.200000	3.400000	2.000000000	101720.00	0	20000107
8	4	6	52.00000	7.200000	7.400000	5.700000	0.000000000	101650.00	0	20000108
9	0	7	55.00000	7.800000	3.200000	-0.700000	0.200000000	102760.00	0	20000109
10	5	4	40.00000	10.200000	2.200000	-3.300000	0.000000000	103470.00	0	20000110

Zestawienie statystyk opisowych zbiorów danych po wykonanych imputacjach do pełnego zbioru danych oraz zbioru danych z wygenerowanymi brakami:

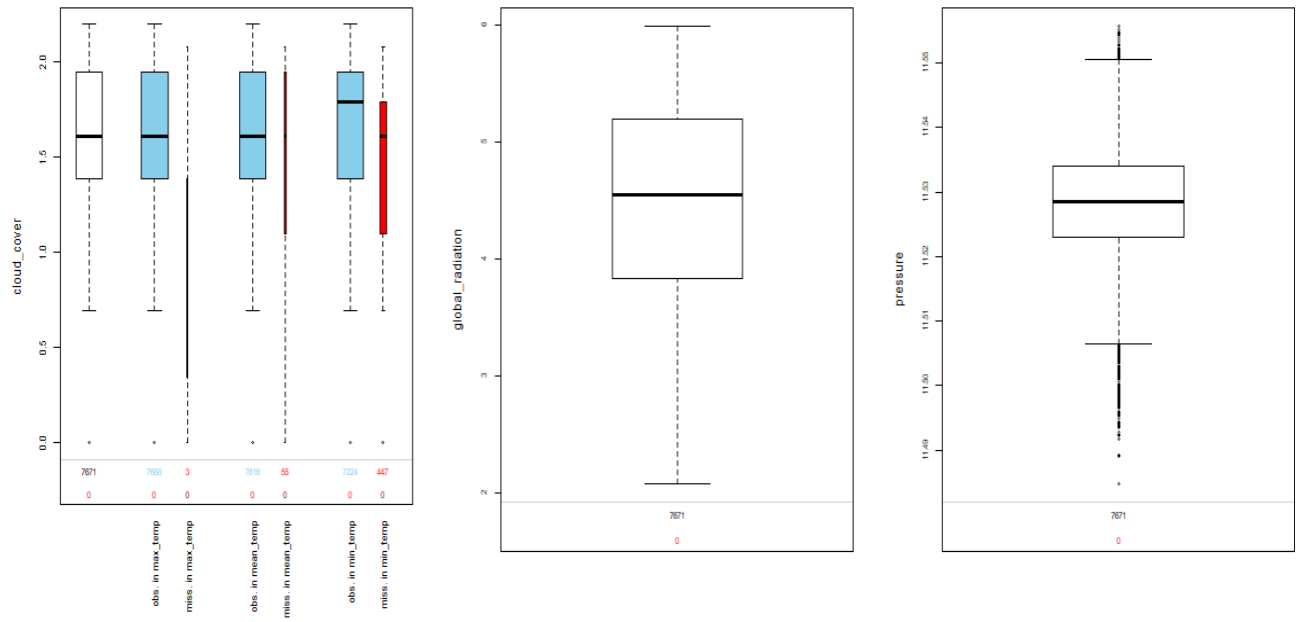
	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth
	Pełen zbiór danych								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	5,06	4,33	118,79	15,81	11,91	8	1,73	101516,7	0,02
Me	5	3,5	95	15,5	11,8	8,2	0	101610	0
s	2,21	3,99	88,53	6,52	5,67	5,24	3,69	1046,43	0,28
Vs	0,44	0,92	0,75	0,41	0,48	0,66	2,13	0,01	13,81
g1	-0,58	0,71	0,67	0,14	0	-0,17	3,79	-0,42	22,12
s(x_śr)	0,98	1,92	8,12	1,64	1,64	1,85	2,81	3,28	1,95
	Zbiór danych z brakami wywołanymi mechanizmem NMAR								
n	5751	5672	5660	5708	5664	5656	5687	5601	5710
x_śr	4,73	4,82	125,44	16,09	11,98	8,02	1,62	101565,95	0,02
Me	5	4,2	103	15,9	11,9	8,2	0	101660	0
s	2,28	4,1	90,8	6,63	5,73	5,28	3,66	1035,35	0,3
Vs	0,48	0,85	0,72	0,41	0,48	0,66	2,26	0,01	13,93
g1	-0,44	0,57	0,59	0,13	0,01	-0,16	4,12	-0,43	21,99
s(x_śr)	1,05	1,87	8,11	1,65	1,65	1,87	2,87	3,25	2,04
	Imputacja metodą k najbliższych sąsiadów (z odległością Gowera)								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	4,94	4,37	118,47	15,86	11,94	8,05	1,78	101535,78	0,02
Me	5	3	94	15,5	11,8	8,2	0,2	101580	0
s	2,12	3,84	85,08	6,28	5,46	5,02	3,37	941,28	0,28
Vs	0,43	0,88	0,72	0,4	0,46	0,62	1,9	0,01	12,81
g1	-0,63	0,77	0,74	0,19	0,02	-0,16	3,96	-0,37	21,87
s(x_śr)	0,95	1,83	7,82	1,58	1,58	1,77	2,53	2,95	1,89
	Imputacja metodą Hot-deck								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	4,87	4,6	122,04	15,84	11,88	7,89	1,75	101543,07	0,02
Me	5	4	98	15,6	11,8	8,1	0	101630	0
s	2,24	4,03	89,96	6,55	5,72	5,3	3,88	1029,6	0,28
Vs	0,46	0,88	0,74	0,41	0,48	0,67	2,21	0,01	13,87
g1	-0,49	0,61	0,63	0,14	0,01	-0,15	3,93	-0,44	22,08
s(x_śr)	1,01	1,88	8,14	1,65	1,66	1,89	2,93	3,23	1,96
	Imputacja metodą Random Forest								
n	7671	7671	7671	7671	7671	7671	7671	7671	7671
x_śr	4,83	4,64	122,53	15,86	11,87	7,91	1,71	101538,9	0,02
Me	5	4	101,83	15,6	11,79	8	0,2	101614,61	0
s	2,11	3,83	87,86	6,51	5,66	5,15	3,45	990,31	0,27
Vs	0,44	0,82	0,72	0,41	0,48	0,65	2,02	0,01	13,07
g1	-0,54	0,65	0,6	0,14	0,01	-0,12	3,96	-0,39	22,14
s(x_śr)	0,96	1,78	7,94	1,63	1,64	1,83	2,64	3,11	1,89

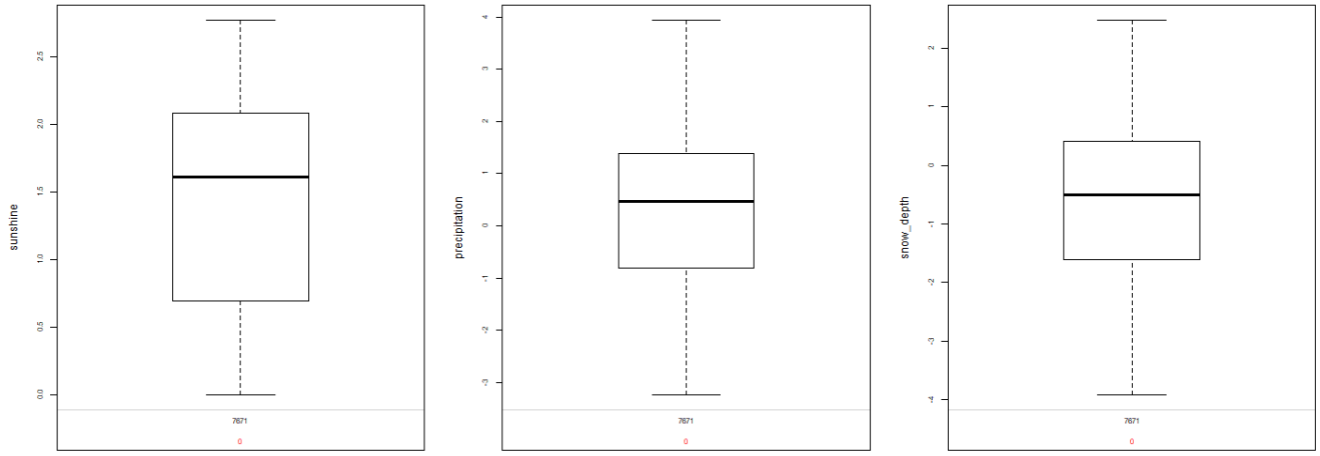
Na podstawie powyższej tabeli można zauważyć, że wszystkie metody imputacji działają poprawnie, ponieważ nie występują żadne braki. Najwięcej brakujących danych zostało wylosowanych dla zmiennej preassure. Kolorem zielonym zostały zaznaczone wartości statystyk opisowych, które są takie same dla pełnego zbioru danych jak i po konkretnej imputacji. W przypadku losowania braków mechanizmem NMAR, również najlepsza okazała się imputacja k najbliższych sąsiadów. Pomimo, że ma mniej identycznych wartości dotyczących statystyk opisowych, to sumarycznie więcej wartości jest zbliżonych do tych dla pełnego zbioru danych. Kolejno najlepiej wypadła imputacja metodą Hot-deck, a na trzecim miejscu lasów losowych.

Wizualizacja rozkładu braków po imputacji metodą k najbliższych sąsiadów (z odległością Gowera):



Wykresy pudełkowe po imputacji metodą k najbliższych sąsiadów (z odległością Gowera):



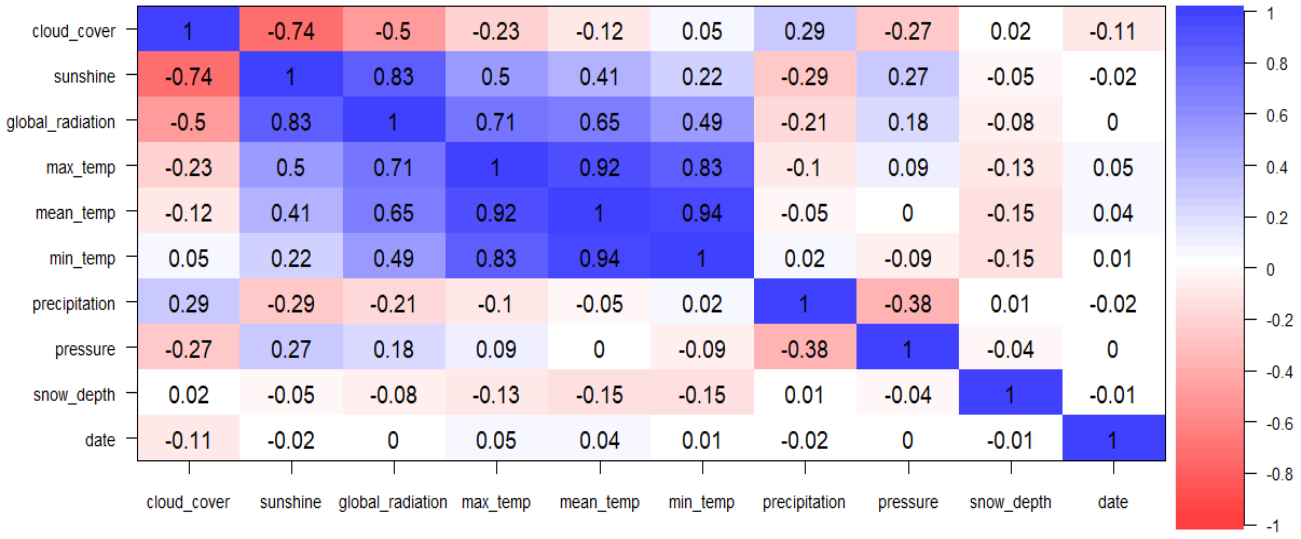


Macierz korelacji zmiennych po wygenerowani braków za pomocą mechanizmu NMAR:

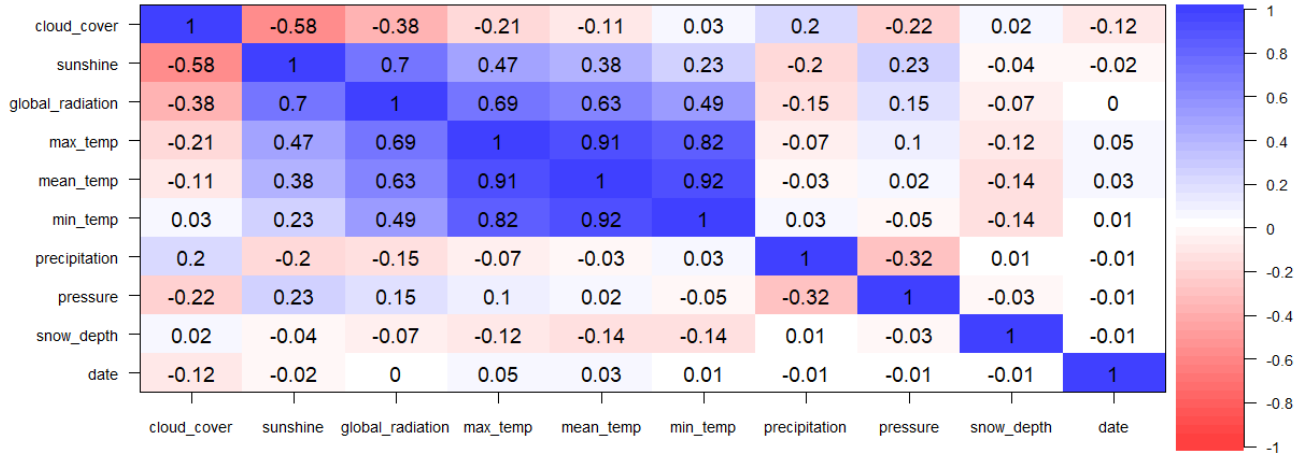
		A	B	C	D	E	F	G	H	I
	r(i)	**cloud_cover**	**sunshine**	**global_radiation**	**max_temp**	**mean_temp**	**min_temp**	**precipitation**	**pressure**	**snow_depth**
A	**cloud_cover**	1	-0,7391	-0,4691	-0,2262	-0,1121	0,0682	0,2811	-0,3084	0,0235
B	**sunshine**	-0,7391	1	0,8325	0,4998	0,4032	0,2173	-0,2685	0,2970	-0,0504
C	**global_radiation**	-0,4691	0,8325	1	0,7031	0,6446	0,4816	-0,2014	0,1830	-0,0671
D	**max_temp**	-0,2262	0,4998	0,7031	1	0,9237	0,8203	-0,0989	0,1164	-0,1269
E	**mean_temp**	-0,1121	0,4032	0,6446	0,9237	1	0,9547	-0,0327	-0,0043	-0,1491
F	**min_temp**	0,0682	0,2173	0,4816	0,8203	0,9547	1	0,0219	-0,0834	-0,1280
G	**precipitation**	0,2811	-0,2685	-0,2014	-0,0989	-0,0327	0,0219	1	-0,3509	0,0060
H	**pressure**	-0,3084	0,2970	0,1830	0,1164	-0,0043	-0,0834	-0,3509	1	-0,0339
I	**snow_depth**	0,0235	-0,0504	-0,0671	-0,1269	-0,1491	-0,1280	0,0060	-0,0339	1

Po wygenerowaniu braków wspomniane wysokie korelacje pomiędzy zmiennymi wciąż są obecne. Korelacja pomiędzy zmiennymi dotyczącymi pomiarów opadu deszczu i śniegu wzrosła jednak wciąż jest bardzo niska, ale w przeciwieństwie do mechanizmu MCAR i MAR ma dodatni znak.

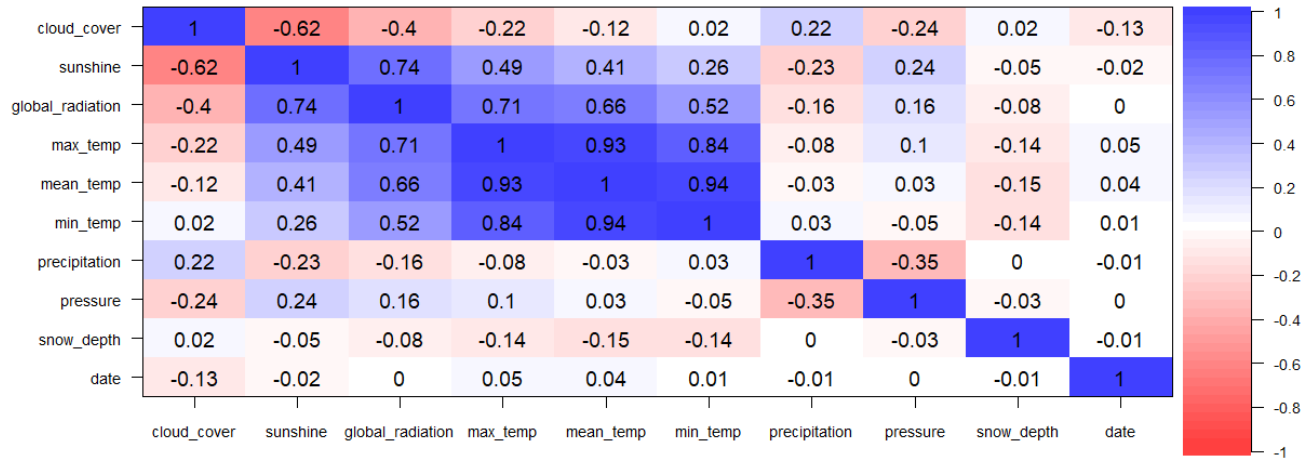
Macierz korelacji po imputacji metodą k najbliższych sąsiadów (z odległością Gowera):



Macierz korelacji po imputacji metodą hot-deck:



Macierz korelacji po imputacji metodą Random forest:



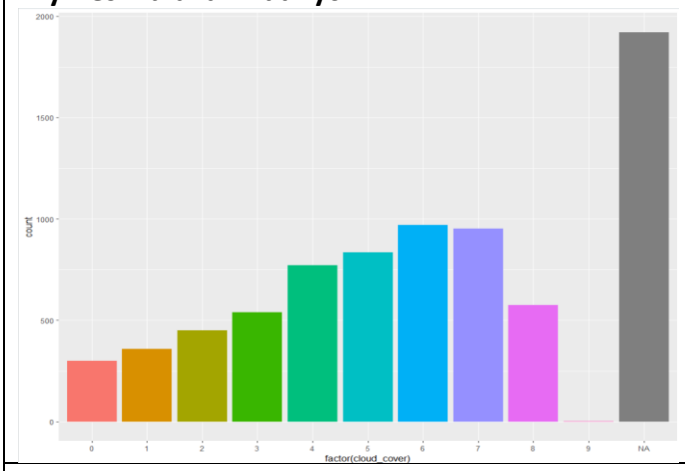
Na podstawie analizy korelogramów można zauważyć, że najbardziej zbliżone wartości korelacji do tych ze zbioru danych bez braków są te, które zostały obliczone na podstawie danych, gdzie braki zimputowano metodą k najbliższych sąsiadów. Korelacje obliczone na podstawie danych, których braki zostały uzupełnione metodą Hot-deck i lasów losowych dają podobne wyniki, jednak dla drugiej metody nieco lepsze.

Wizualizacja wybranych zmiennych przed imputacją braków wywołanych mechanizmem NMAR i po imputacji

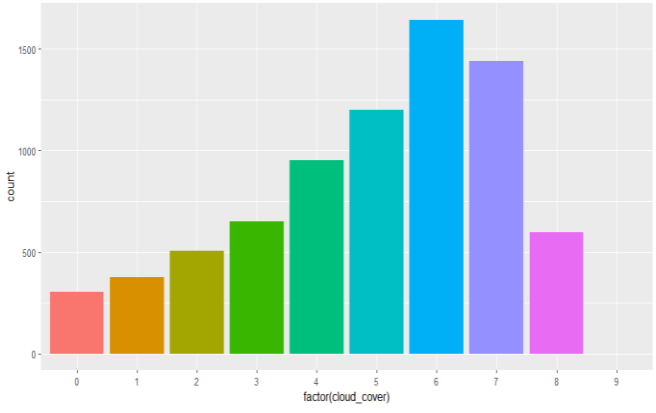
We wszystkich wizualizacjach pełnego zbioru danych, zbioru danych z brakami oraz zbiorów danych po imputacjach będą porównywane te same zmienne w celu zaobserwowania różnic

Zmienna cloud_cover:

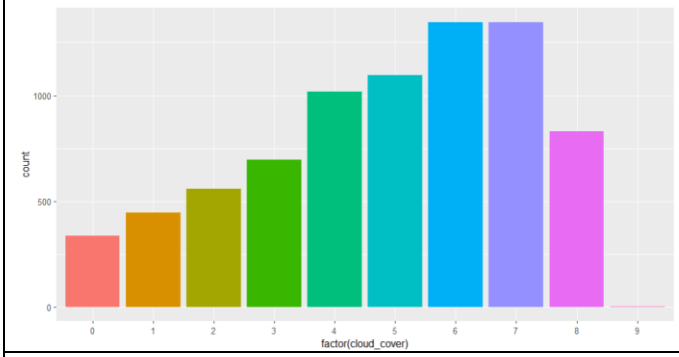
Wykres z brakami danych:



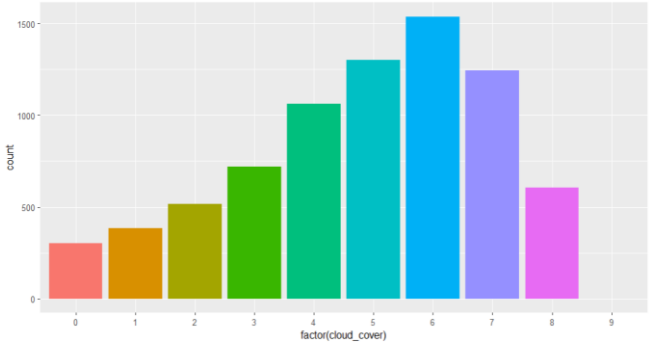
Imputacja metodą k najbliższych sąsiadów (z odległością Gowera):



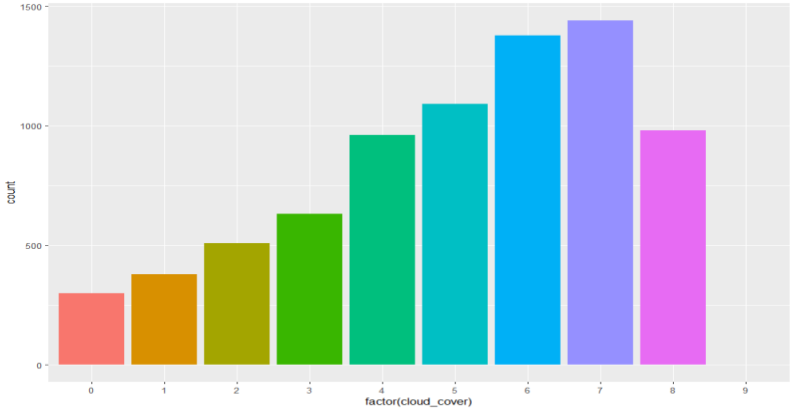
Imputacja metodą Hot-deck:



Imputacja metodą Random Forest:



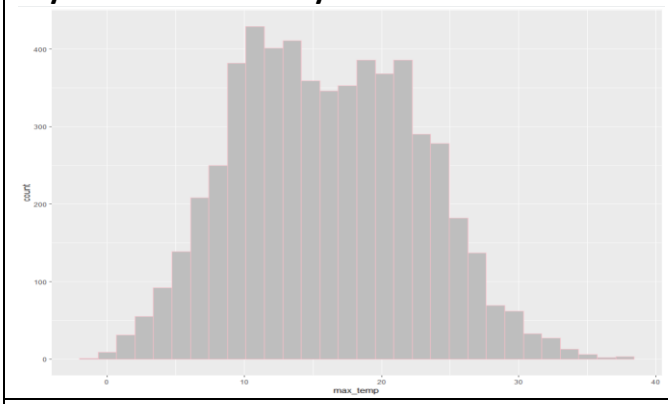
Wykres bez braków danych :



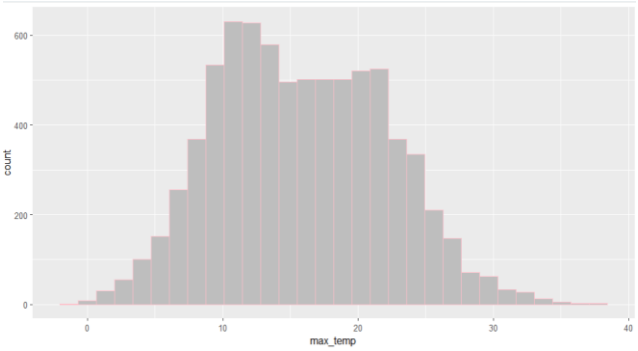
Analizując wykresy zmiennej cloud_cover przed i po imputacji braków danych można zauważyć, że patrząc na samą wizualizację zmiennych najbardziej zbliżony wykres do tego wykonanego na pełnym zbiorze danych to ten, wykonany na danych po imputacji braków metodą Hot-deck. Kolejno najbardziej zbliżony jest wykres danych po imputacji braków metodą k najbliższych sąsiadów i na końcu metodą lasów losowych. Największe różnice można zauważyć dla wartości 6 i 7 zmiennej, pomiędzy wykresami.

Zmienna max_temp:

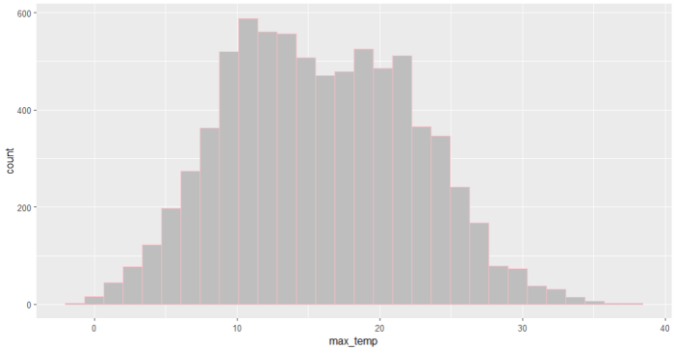
Wykres z brakami danych:



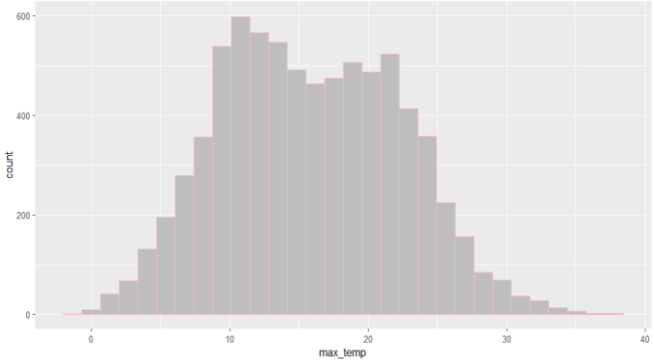
Imputacja metodą k najbliższych sąsiadów (z odległością Gowera):



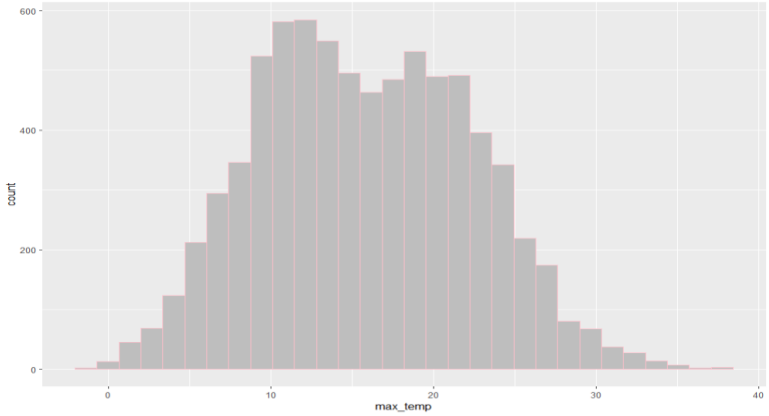
Imputacja metodą Hot-deck:



Imputacja metodą Random Forest:



Wykres bez braków danych :



Analizując wykresy zmiennej max_temp przed i po imputacji braków danych można zauważyć, że najbardziej obiegają te dla danych, gdzie braki zostały zimputowane za pomocą metody Hot-deck i lasów losowych. Najbliższy wykresowi wykonanemu na pełnym zbiorze danych jest wykres po imputacji braków danych metodą k-najbliższych sąsiadów. Podsumowując wszystkie wykresy stosunkowo dobrze odwzorowują wykres wykonany na zbiorze danych bez braków.

Wnioski:

Przy mechanizmie losowania braków NMAR wszystkie metody okazały się odpowiednie, jednak najlepszą niezmiennie pozostaje metoda k najbliższych sąsiadów, a kolejno lasów i na trzecim miejscu hot-deck.