Case Studies: Health Data

Muhammad Raafey Tariq Farrukh Ahmed Aymane Hachcham **Amirreza Khamehchin Khiabani**

Technische Universität Dortmund

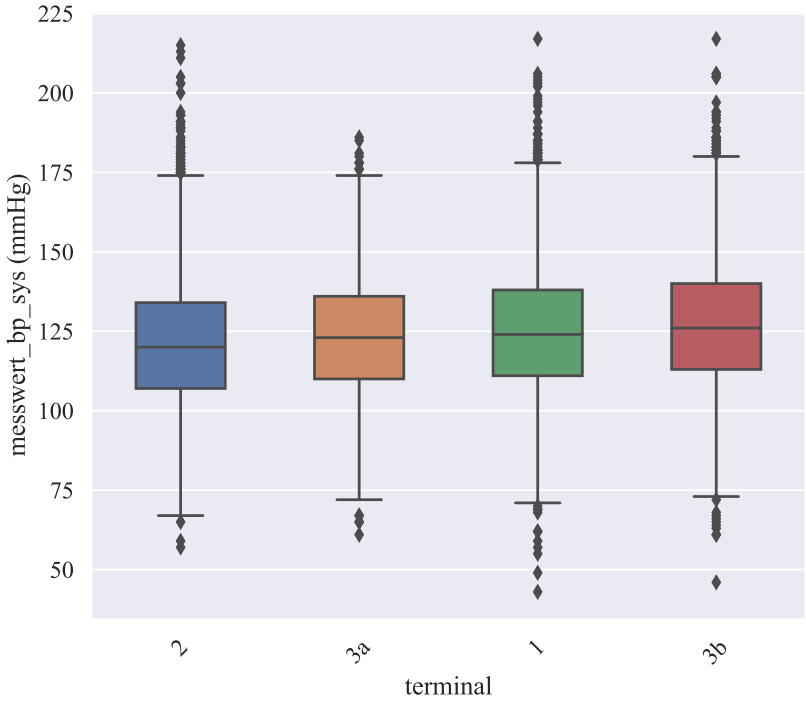
May 21, 2023

Adjusted variable terminal

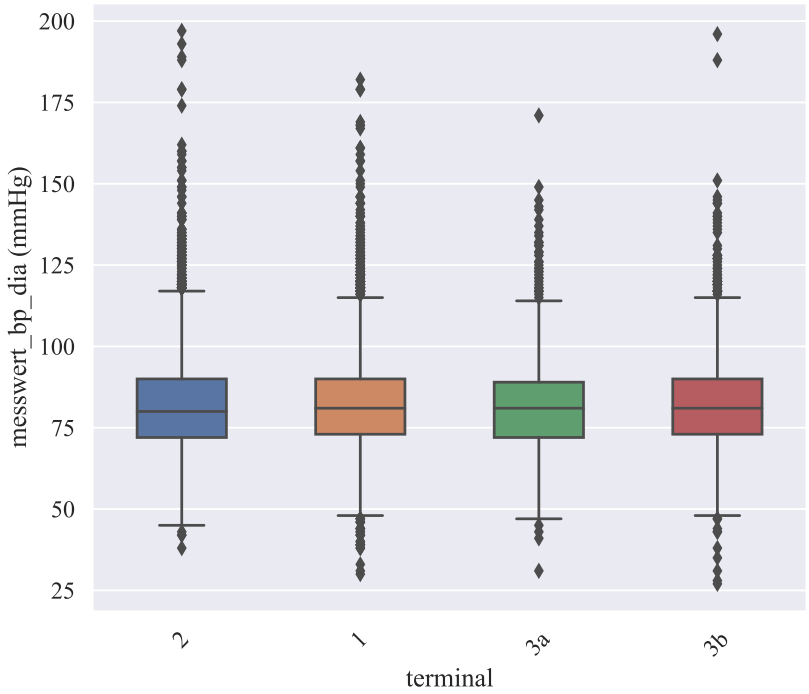
Split terminal 3 into two subgroups 3a and 3b after measurement device was changed

Terminal	Count
1	5699
2	4655
3a	1397
3b	3080

Terminal on the blood pressures



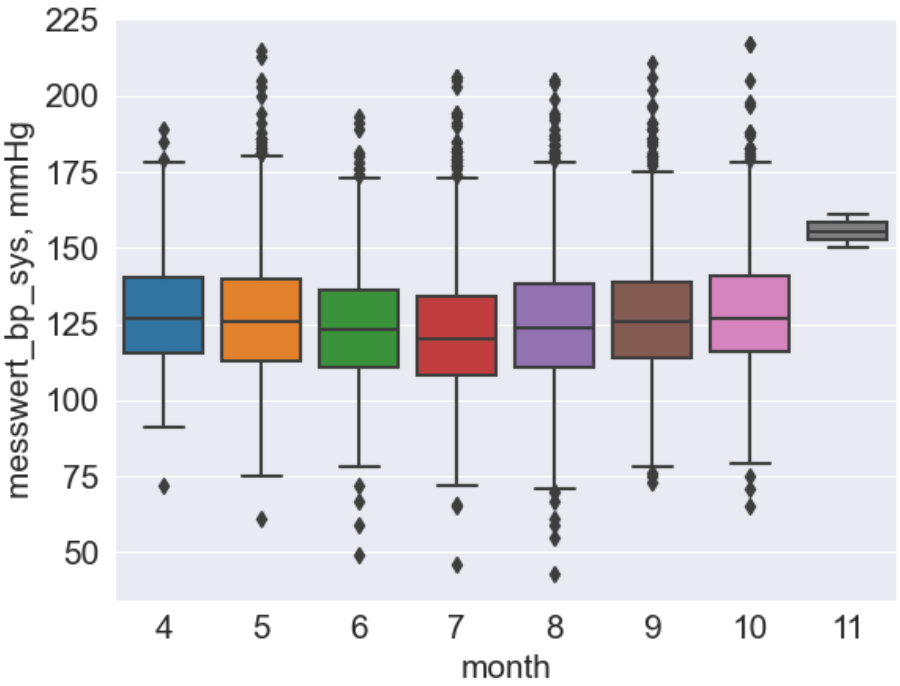
(a) Boxplot for the systolic blood pressure by terminal



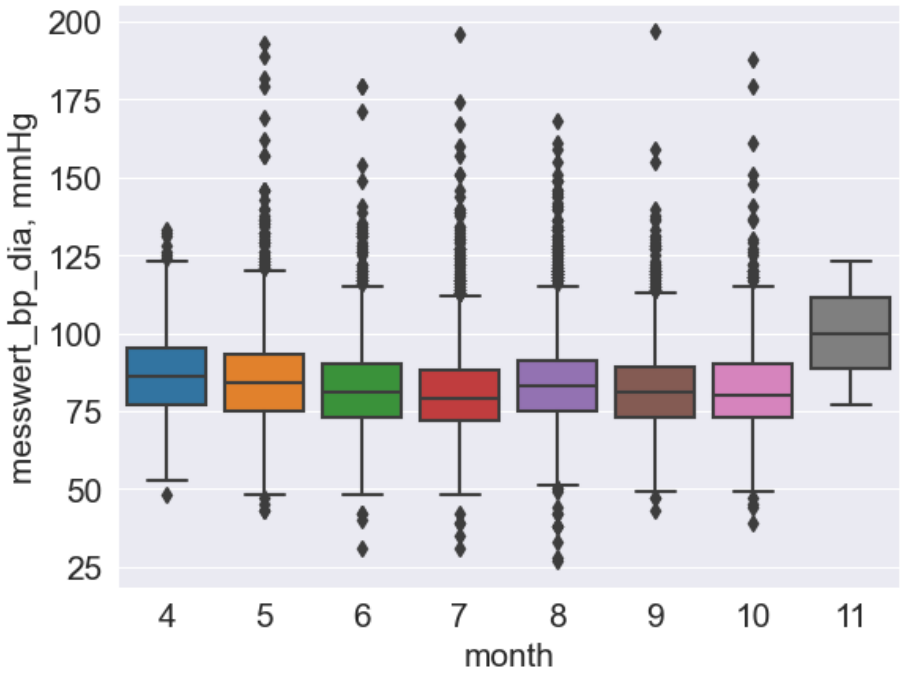
(b) Boxplot for the diastolic blood pressure by terminal

Figure: Effect of terminals on different blood pressure measurements

Time Effect on the blood pressures



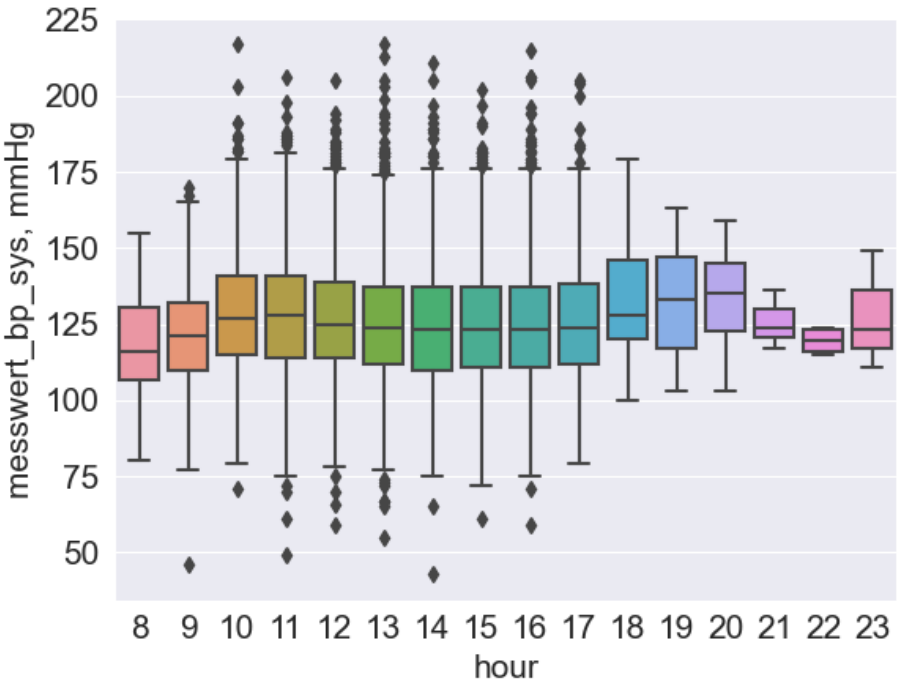
(a) Boxplot for the systolic blood pressure by month



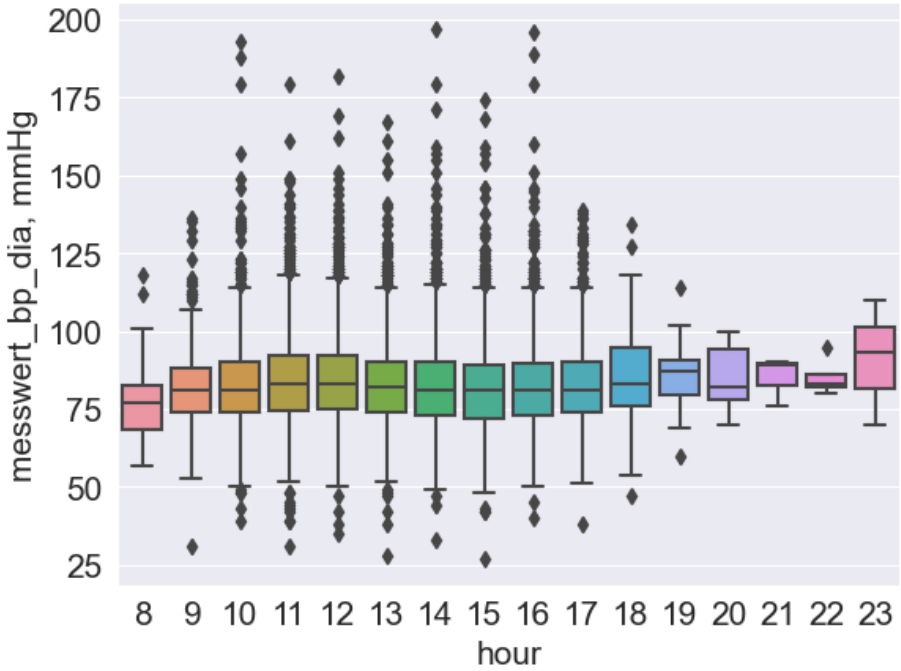
(b) Boxplot for the diastolic blood pressure by month

Figure: Effect of the month of the different blood pressures

Time Effect on the blood pressures



(a) Boxplot for the systolic blood pressure by hour



(b) Boxplot for the diastolic blood pressure by hour

Figure: Effect of the hour of the different blood pressures

Homogeneity of the blood pressure across all months

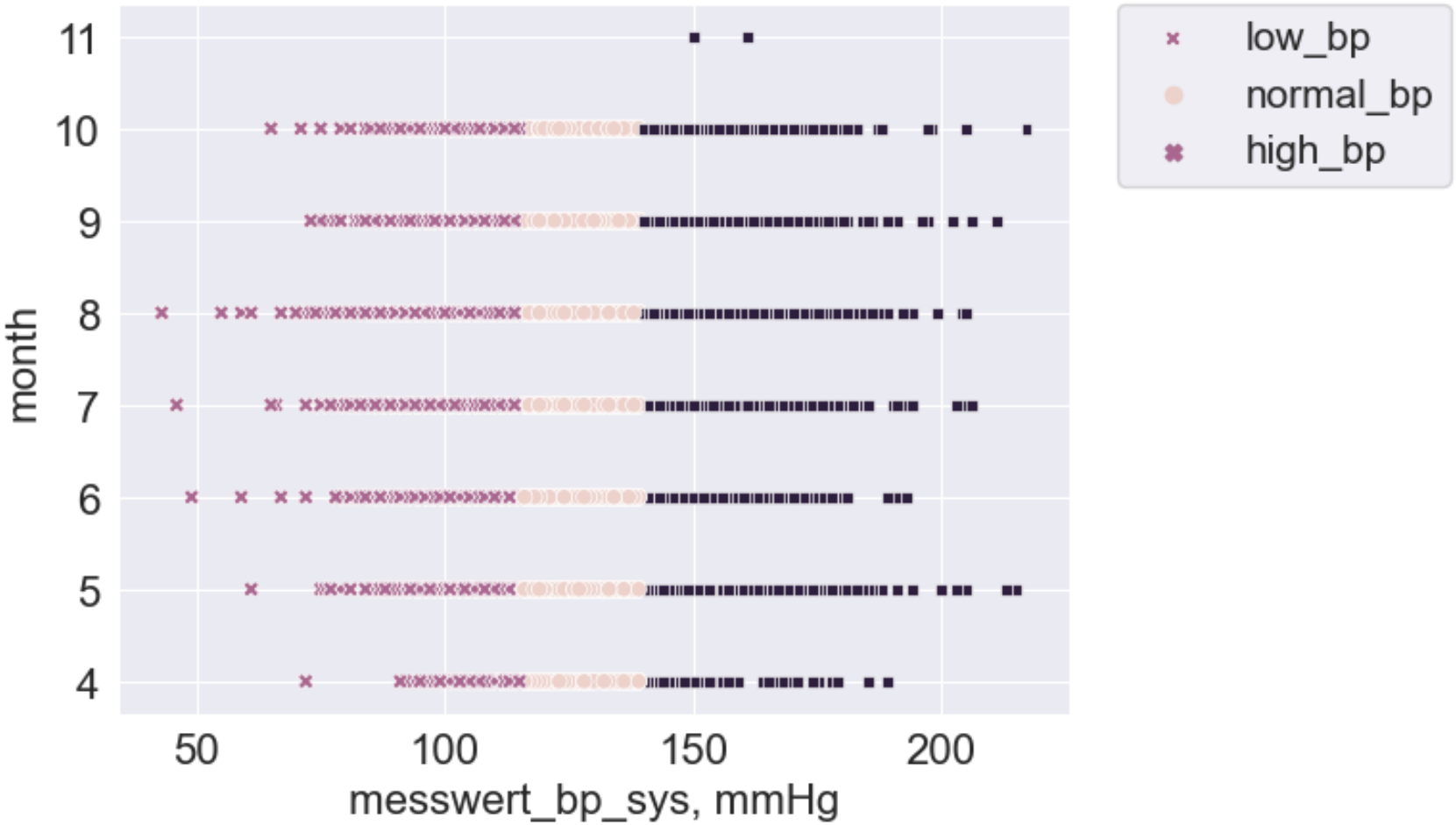


Figure: Cluster distributions for the systolic bp across 11 months

Homogeneity of the blood pressure across all months

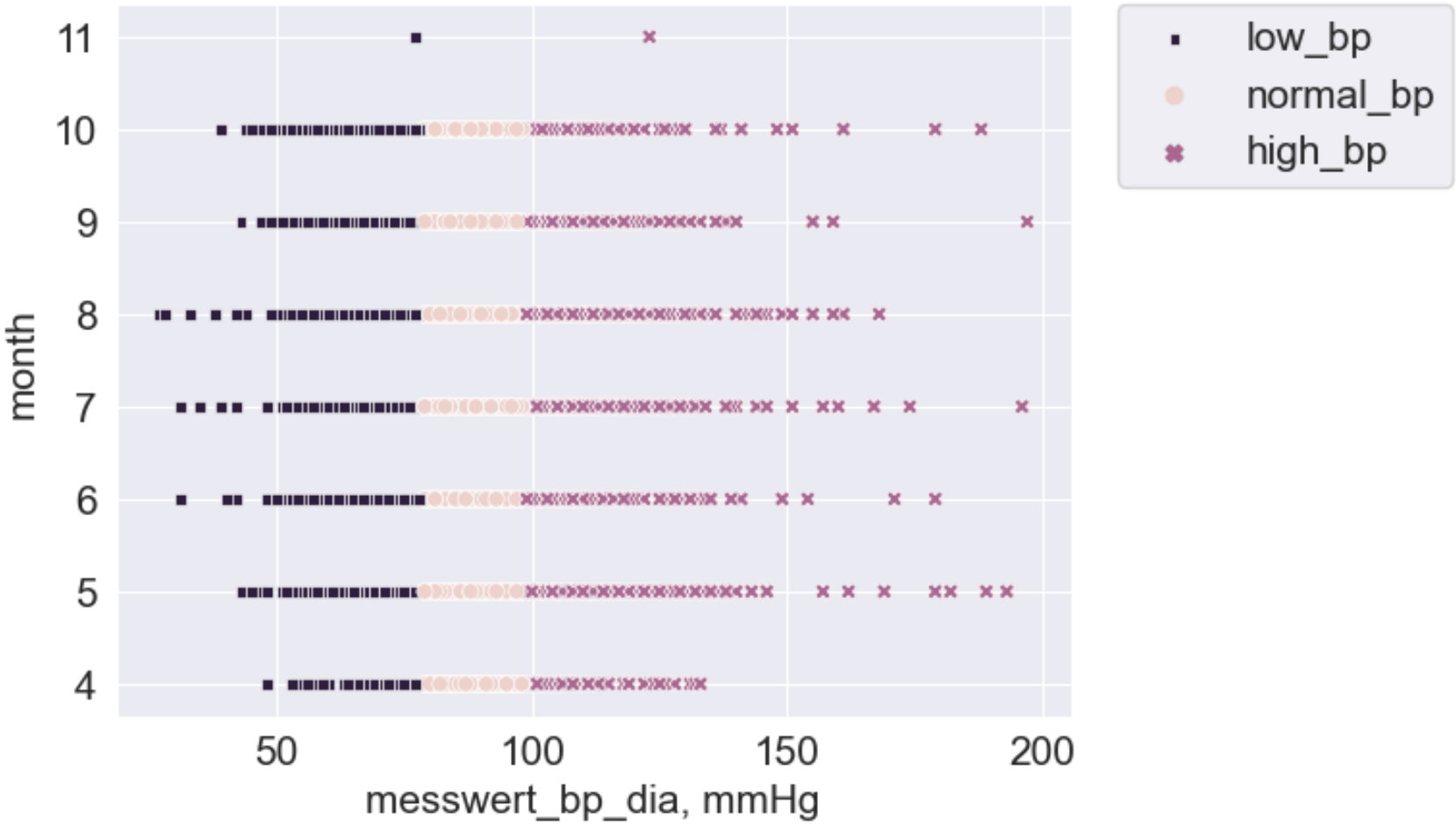
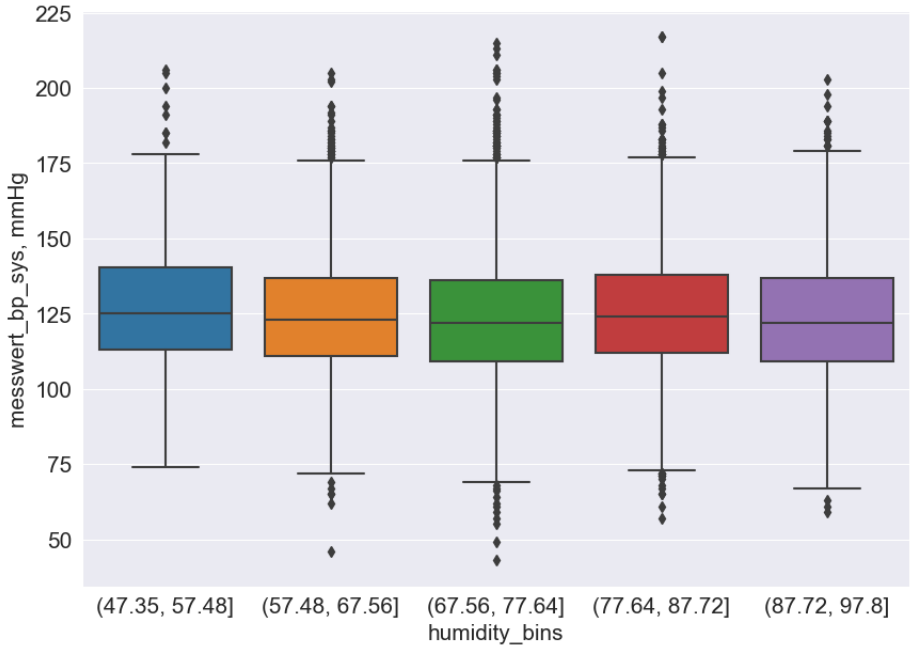


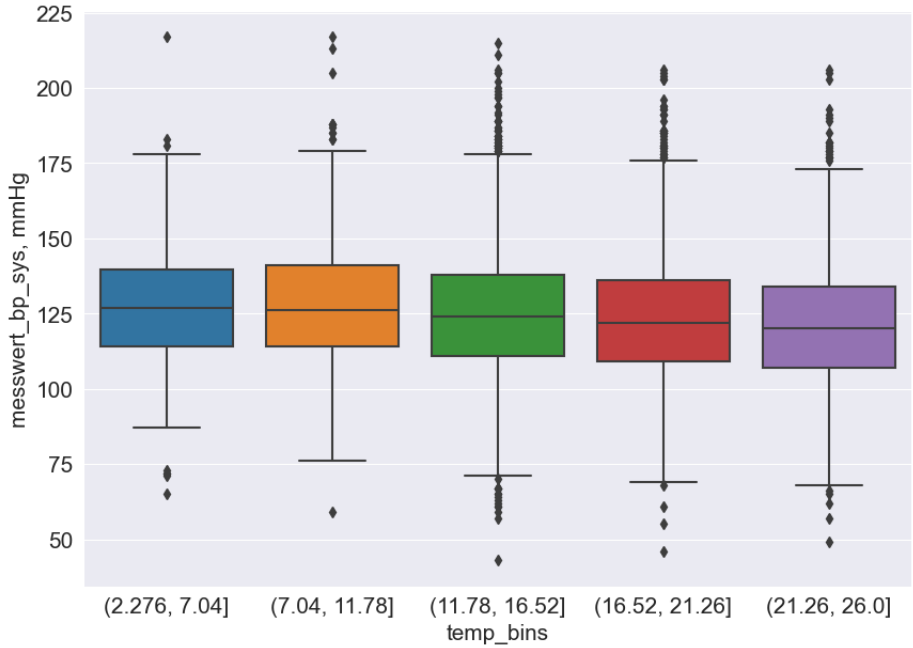
Figure: Cluster distributions for the diastolic bp across 11 months

Weather condition effects on the blood pressures

Collected temperature and humidity stats from visualcrossing.com for the corresponding dates



(a) Effect of humidity



(b) Effect of temperature

Results of the linear model for the systolic bp.

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	45.0035	2.699	16.677	0.000	39.714	50.293
terminal_2	-3.2132	0.311	-10.336	0.000	-3.823	-2.604
terminal_3a	0.3850	0.510	0.755	0.450	-0.614	1.384
terminal_3b	1.8194	0.364	4.998	0.000	1.106	2.533
bundesland_Kärnten	0.9130	1.851	0.493	0.622	-2.716	4.542
bundesland_Niederösterreich	0.5294	1.653	0.320	0.749	-2.710	3.769
bundesland_Oberösterreich	1.7813	1.826	0.976	0.329	-1.798	5.361
bundesland_Salzburg	-1.2314	2.279	-0.540	0.589	-5.699	3.236
bundesland_Steiermark	0.5770	1.477	0.391	0.696	-2.319	3.473
bundesland_Tirol	-2.8701	2.358	-1.217	0.223	-7.491	1.751
bundesland_Vorarlberg	-3.0991	3.116	-0.995	0.320	-9.207	3.009
bundesland_Wien	-0.2240	1.635	-0.137	0.891	-3.428	2.980
bundesland_not_applicable	0.7396	1.776	0.416	0.677	-2.741	4.220
befinden_2	0.5334	0.291	1.830	0.067	-0.038	1.105
befinden_3	0.2204	0.415	0.531	0.596	-0.594	1.035
befinden_4	-0.8226	1.176	-0.699	0.484	-3.128	1.483
befinden_5	-0.7040	2.124	-0.331	0.740	-4.868	3.460
geschlecht_m	1.6358	0.264	6.194	0.000	1.118	2.154
raucher_True	-0.7735	0.359	-2.153	0.031	-1.478	-0.069
blutzucker_bekannt_True	0.9813	0.364	2.695	0.007	0.268	1.695
cholesterin_bekannt_True	-0.6195	0.338	-1.834	0.067	-1.282	0.043
in_behandlung_True	6.0776	0.412	14.754	0.000	5.270	6.885
messwert_bp_dia	0.8446	0.009	90.419	0.000	0.826	0.863
age	0.2363	0.009	26.107	0.000	0.219	0.254
month	0.7599	0.097	7.817	0.000	0.569	0.950
hour	-0.1019	0.058	-1.749	0.080	-0.216	0.012
day	0.1069	0.015	6.996	0.000	0.077	0.137
temp	-0.1802	0.187	-0.964	0.335	-0.547	0.186
humidity	-0.0406	0.019	-2.194	0.028	-0.077	-0.004
temp_min	-0.4771	0.120	-3.976	0.000	-0.712	-0.242
temp_max	0.1454	0.092	1.577	0.115	-0.035	0.326

Results of the linear model for the diastolic bp.

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	26.0262	2.136	12.187	0.000	21.840	30.212
terminal_2	1.8276	0.245	7.456	0.000	1.347	2.308
terminal_3a	-1.2640	0.401	-3.154	0.002	-2.050	-0.478
terminal_3b	-0.3386	0.287	-1.181	0.237	-0.901	0.223
bundesland_Kärnten	-1.2178	1.456	-0.836	0.403	-4.072	1.636
bundesland_Niederösterreich	-0.9734	1.300	-0.749	0.454	-3.522	1.575
bundesland_Oberösterreich	-2.4861	1.436	-1.731	0.083	-5.301	0.329
bundesland_Salzburg	-1.6790	1.793	-0.937	0.349	-5.193	1.835
bundesland_Steiermark	-1.0891	1.162	-0.937	0.349	-3.367	1.189
bundesland_Tirol	-0.2784	1.855	-0.150	0.881	-3.914	3.357
bundesland_Vorarlberg	0.2043	2.451	0.083	0.934	-4.600	5.009
bundesland_Wien	-0.1171	1.286	-0.091	0.927	-2.637	2.403
bundesland_not_applicable	-1.7818	1.397	-1.276	0.202	-4.519	0.956
befinden_2	-0.6128	0.229	-2.673	0.008	-1.062	-0.163
befinden_3	-0.5423	0.327	-1.660	0.097	-1.183	0.098
befinden_4	0.6051	0.925	0.654	0.513	-1.208	2.419
befinden_5	4.6722	1.670	2.797	0.005	1.398	7.946
geschlecht_m	1.2121	0.208	5.834	0.000	0.805	1.619
raucher_True	0.6494	0.283	2.298	0.022	0.096	1.203
blutzucker_bekannt_True	-0.1925	0.286	-0.672	0.502	-0.754	0.369
cholesterin_bekannt_True	0.1666	0.266	0.627	0.531	-0.354	0.688
in_behandlung_True	-1.6254	0.327	-4.971	0.000	-2.266	-0.984
messwert_bp_sys	0.5225	0.006	90.419	0.000	0.511	0.534
age	-0.0802	0.007	-10.974	0.000	-0.094	-0.066
month	-0.7998	0.076	-10.484	0.000	-0.949	-0.650
hour	-0.0737	0.046	-1.607	0.108	-0.164	0.016
day	-0.0627	0.012	-5.209	0.000	-0.086	-0.039
temp	0.4083	0.147	2.776	0.006	0.120	0.697
humidity	0.0420	0.015	2.881	0.004	0.013	0.071
temp_min	0.0119	0.094	0.126	0.900	-0.173	0.197
temp_max	-0.2857	0.073	-3.941	0.000	-0.428	-0.144

Results for LM: Diastolic as target

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	LM (Base)	109.906435	108.071011	0.456546	0.454677	0.455392	0.451967
1	LM (Best Subset)	108.053036	110.150122	0.454768	0.455342	0.453293	0.454711

Table: Results before adding new variables

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	LM (Base)	106.417234	109.579047	0.468764	0.459289	0.467224	0.455618
1	LM (Best Subset)	106.417234	109.579047	0.468764	0.459289	0.467224	0.455618

Table: Results after adding new variables

Results for LM: Systolic as target

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	LM (Base)	177.574175	173.376931	0.522072	0.527149	0.521057	0.524799
1	LM (Best Subset)	173.320480	177.601697	0.527303	0.521998	0.525276	0.521121

Table: Results before adding new variables

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	LM (Base)	172.011759	171.426840	0.537434	0.531401	0.536093	0.528220
1	LM (Best Subset)	172.593103	171.291016	0.535870	0.531772	0.534615	0.528807

Table: Results after adding new variables

Results for RT: Diastolic as target

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	Tree (Base)	0.490000	217.600000	1.000000	-0.100000	1.000000	-0.110000
1	Tree (Fine-tuned)	100.350000	105.500000	0.500000	0.460000	0.500000	0.460000
2	RF (Base)	15.950000	114.530000	0.920000	0.420000	0.920000	0.420000
3	RF (Fine-tuned)	97.310000	103.100000	0.520000	0.480000	0.520000	0.470000

Table: Results before adding new variables

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	Tree (Base)	0.200000	207.990000	1.000000	-0.100000	1.000000	-0.110000
1	Tree (Fine-tuned)	103.240000	106.830000	0.500000	0.430000	0.500000	0.430000
2	RF (Base)	15.090000	105.740000	0.930000	0.440000	0.930000	0.440000
3	RF (Fine-tuned)	99.900000	103.490000	0.520000	0.450000	0.510000	0.450000

Table: Results after adding new variables

Results for RT: Systolic as target

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	Tree (Base)	1.240000	358.490000	1.000000	0.040000	1.000000	0.040000
1	Tree (Fine-tuned)	162.760000	186.530000	0.560000	0.500000	0.560000	0.500000
2	RF (Base)	25.890000	190.670000	0.930000	0.490000	0.930000	0.490000
3	RF (Fine-tuned)	161.510000	177.670000	0.560000	0.520000	0.560000	0.520000

Table: Results before adding new variables

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	Tree (Base)	0.030000	347.170000	1.000000	0.050000	1.000000	0.040000
1	Tree (Fine-tuned)	164.050000	184.730000	0.560000	0.490000	0.560000	0.490000
2	RF (Base)	23.980000	175.470000	0.940000	0.520000	0.940000	0.520000
3	RF (Fine-tuned)	161.560000	177.680000	0.570000	0.510000	0.560000	0.510000

Table: Results after adding new variables

Summary

- Added new variables: *day*, *month*, *hour*, *temp*, *temp_min*, *temp_max*, *and humidity*.
- The effects of the new variables on the target have been analyzed.
- Results of the newly fitted models are compared to the old models.
- Slight improvements in the mean squared errors for both Linear Models and Regression Trees.

Follow up questions:

1. Should the new variables be kept even if the improvement was a minor one?
2. Should the variable for month, time, day be a categorical or numerical variable?

Thank you!

Questions?