Analyse

April 12, 2024

1 Erste Analyse

1.1 Get some baseline facts

```
1805
```

```
5895
```

1.2 How many follow-up assessments over the time?

Number of monthly active users in baseline

YYYY-MM	
2020-07	89
2020-08	566
2020-09	74
2020-10	204
2020-11	132
2020-12	137
2021-01	88
2021-02	41
2021-03	79
2021-04	57
2021-05	45
2021-06	34
2021-07	18
2021-08	7
2021-09	7
2021-10	64
2021-11	39
2021-12	22
2022-01	24
2022-02	17
2022-03	4
2022-04	2
2022-05	11
2022-06	7
2022-07	5

```
2022-08      4
2022-09      3
2022-10      9
2022-11      4
2023-01      2
2023-02      2
2023-03      1
2023-04      1
2023-05      2
2023-08      2
2023-09      2
Name: user_id, dtype: int64
```

n assessments per month in follow up

YYYY-MM	
2020-07	1
2020-08	397
2020-09	427
2020-10	395
2020-11	438
2020-12	445
2021-01	409
2021-02	348
2021-03	347
2021-04	309
2021-05	297
2021-06	209
2021-07	156
2021-08	139
2021-09	108
2021-10	129
2021-11	171
2021-12	134
2022-01	141
2022-02	68
2022-03	102
2022-04	56
2022-05	150
2022-06	85
2022-07	64
2022-08	79
2022-09	58
2022-10	48
2022-11	35
2022-12	27
2023-01	28

```
2023-02     21
2023-03     19
2023-04     19
2023-05     12
2023-06      6
2023-07      6
2023-08      6
2023-09      2
2023-10      4
Name: count, dtype: int64
```

Number of monthly active users in follow up

```
YYYY-MM
2020-07      1
2020-08    308
2020-09    257
2020-10    249
2020-11    255
2020-12    232
2021-01    236
2021-02    199
2021-03    186
2021-04    174
2021-05    161
2021-06    127
2021-07     93
2021-08     70
2021-09     64
2021-10     80
2021-11     87
2021-12     76
2022-01     77
2022-02     56
2022-03     45
2022-04     28
2022-05     42
2022-06     46
2022-07     37
2022-08     39
2022-09     36
2022-10     28
2022-11     24
2022-12     17
2023-01     17
2023-02     13
2023-03     12
```

```
2023-04     14
2023-05      8
2023-06      4
2023-07      4
2023-08      3
2023-09      2
2023-10      2
Name: user_id, dtype: int64
```

1.3 Welche Domänen (alle im Fragebogen) wurden über die Zeit wie häufig befüllt und ausgefüllt im Zeitverlauf

Zwischen den Domänen sind keine großen Unterschiede im Zeitverlauf. Wenn es in einem Monat viele aktive Nutzer gab, dann auch in den jeweiligen Domänen. Die unten stehende Tabelle gibt die Anzahl der vorhandenen Werte in diesem Jahr und Monat (YYYY-MM) für alle Variablen an.

YYYY-MM	questionnaire_id	user_id	created_at	sensordata_apps	\
2020-07	89	89	89	9	
2020-08	566	566	566	12	
2020-09	74	74	74	3	
2020-10	204	204	204	12	
2020-11	132	132	132	6	
2020-12	137	137	137	11	
2021-01	88	88	88	5	
2021-02	41	41	41	3	
2021-03	79	79	79	5	
2021-04	57	57	57	4	
2021-05	45	45	45	3	
2021-06	34	34	34	1	
2021-07	18	18	18	5	
2021-08	7	7	7	0	
2021-09	7	7	7	0	
2021-10	64	64	64	4	
2021-11	39	39	39	7	
2021-12	22	22	22	1	
2022-01	24	24	24	3	
2022-02	17	17	17	1	
2022-03	4	4	4	0	
2022-04	2	2	2	0	
2022-05	11	11	11	1	
2022-06	7	7	7	0	
2022-07	5	5	5	0	
2022-08	4	4	4	1	
2022-09	3	3	3	1	
2022-10	9	9	9	0	

2022-11	4	4	4	1
2023-01	2	2	2	0
2023-02	2	2	2	1
2023-03	1	1	1	0
2023-04	1	1	1	1
2023-05	2	2	2	0
2023-08	2	2	2	0
2023-09	2	2	2	0

YYYY-MM	sensordata_beginTime	sensordata_collected_at	sensordata_endTime	\
2020-07	9	68	9	
2020-08	12	485	12	
2020-09	3	62	3	
2020-10	12	163	12	
2020-11	6	107	6	
2020-12	11	114	11	
2021-01	5	74	5	
2021-02	3	26	3	
2021-03	5	60	5	
2021-04	4	42	4	
2021-05	3	36	3	
2021-06	1	27	1	
2021-07	5	13	5	
2021-08	0	4	0	
2021-09	0	6	0	
2021-10	4	51	4	
2021-11	7	34	7	
2021-12	1	18	1	
2022-01	3	19	3	
2022-02	1	12	1	
2022-03	0	3	0	
2022-04	0	2	0	
2022-05	1	6	1	
2022-06	0	5	0	
2022-07	0	4	0	
2022-08	1	3	1	
2022-09	1	3	1	
2022-10	0	7	0	
2022-11	1	4	1	
2023-01	0	1	0	
2023-02	1	1	1	
2023-03	0	1	0	
2023-04	1	1	1	
2023-05	0	1	0	
2023-08	0	2	0	
2023-09	0	2	0	

YYYY-MM	sensordata_name	sensordata_sleepTimes	sensordata_top5Apps	…	\
				…	
2020-07	68	9	9	…	
2020-08	485	12	12	…	
2020-09	62	3	3	…	
2020-10	163	12	12	…	
2020-11	107	6	6	…	
2020-12	114	11	11	…	
2021-01	74	5	5	…	
2021-02	26	3	3	…	
2021-03	60	5	5	…	
2021-04	42	4	4	…	
2021-05	36	3	3	…	
2021-06	27	1	1	…	
2021-07	13	5	5	…	
2021-08	4	0	0	…	
2021-09	6	0	0	…	
2021-10	51	4	4	…	
2021-11	34	7	7	…	
2021-12	18	1	1	…	
2022-01	19	3	3	…	
2022-02	12	1	1	…	
2022-03	3	0	0	…	
2022-04	2	0	0	…	
2022-05	6	1	1	…	
2022-06	5	0	0	…	
2022-07	4	0	0	…	
2022-08	3	1	1	…	
2022-09	3	1	1	…	
2022-10	7	0	0	…	
2022-11	4	1	1	…	
2023-01	1	0	0	…	
2023-02	1	1	1	…	
2023-03	1	0	0	…	
2023-04	1	1	1	…	
2023-05	1	0	0	…	
2023-08	2	0	0	…	
2023-09	2	0	0	…	

YYYY-MM	pain1	pain2	name	device	os	sensordata	sensordata_altitude	\
2020-07	89	89	89	89	89	0	59	
2020-08	566	566	566	566	566	0	473	
2020-09	74	74	74	74	74	0	59	
2020-10	204	204	204	204	204	0	151	
2020-11	132	132	132	132	132	0	101	

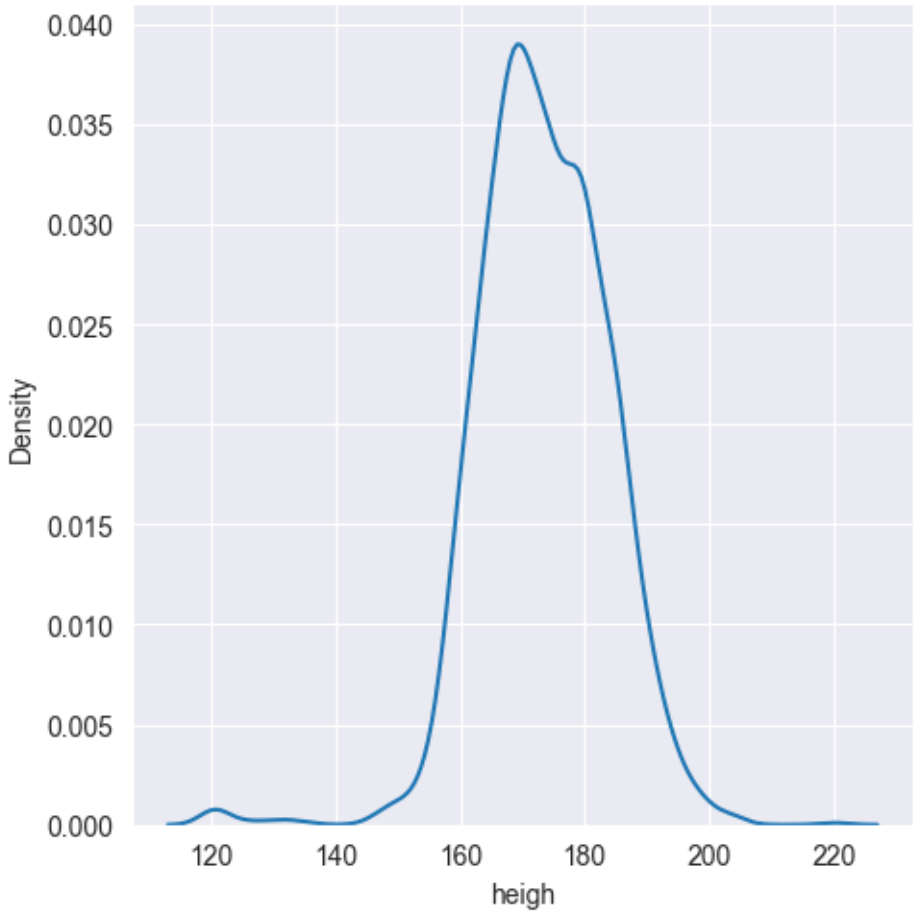
2020-12	137	137	137	137	137	0	103
2021-01	88	88	88	88	88	0	69
2021-02	41	41	41	41	41	0	23
2021-03	79	79	79	79	79	0	55
2021-04	57	57	57	57	57	0	38
2021-05	45	45	45	45	45	0	33
2021-06	34	34	34	34	34	0	26
2021-07	18	18	18	18	18	0	8
2021-08	7	7	7	7	7	0	4
2021-09	7	7	7	7	7	0	6
2021-10	64	64	64	64	64	0	47
2021-11	39	39	39	39	39	0	27
2021-12	22	22	22	22	22	0	17
2022-01	24	24	24	24	24	0	16
2022-02	17	17	17	17	17	0	11
2022-03	4	4	4	4	4	0	3
2022-04	2	2	2	2	2	0	2
2022-05	11	11	11	11	11	0	5
2022-06	7	7	7	7	7	0	5
2022-07	5	5	5	5	5	0	4
2022-08	4	4	4	4	4	0	2
2022-09	3	3	3	3	3	0	2
2022-10	9	9	9	9	9	0	7
2022-11	4	4	4	4	4	0	3
2023-01	2	2	2	2	2	0	1
2023-02	2	2	2	2	2	0	0
2023-03	1	1	1	1	1	0	1
2023-04	1	1	1	1	1	0	0
2023-05	2	2	2	2	2	0	1
2023-08	2	2	2	2	2	0	2
2023-09	2	2	2	2	2	0	2

YYYY-MM	sensordata_latitude	sensordata_longitude	YYYY-MM
2020-07	59	59	89
2020-08	473	473	566
2020-09	59	59	74
2020-10	151	151	204
2020-11	101	101	132
2020-12	103	103	137
2021-01	69	69	88
2021-02	23	23	41
2021-03	55	55	79
2021-04	38	38	57
2021-05	33	33	45
2021-06	26	26	34
2021-07	8	8	18

2021-08	4	4	7
2021-09	6	6	7
2021-10	47	47	64
2021-11	27	27	39
2021-12	17	17	22
2022-01	16	16	24
2022-02	11	11	17
2022-03	3	3	4
2022-04	2	2	2
2022-05	5	5	11
2022-06	5	5	7
2022-07	4	4	5
2022-08	2	2	4
2022-09	2	2	3
2022-10	7	7	9
2022-11	3	3	4
2023-01	1	1	2
2023-02	0	0	2
2023-03	1	1	1
2023-04	0	0	1
2023-05	1	1	2
2023-08	2	2	2
2023-09	2	2	2

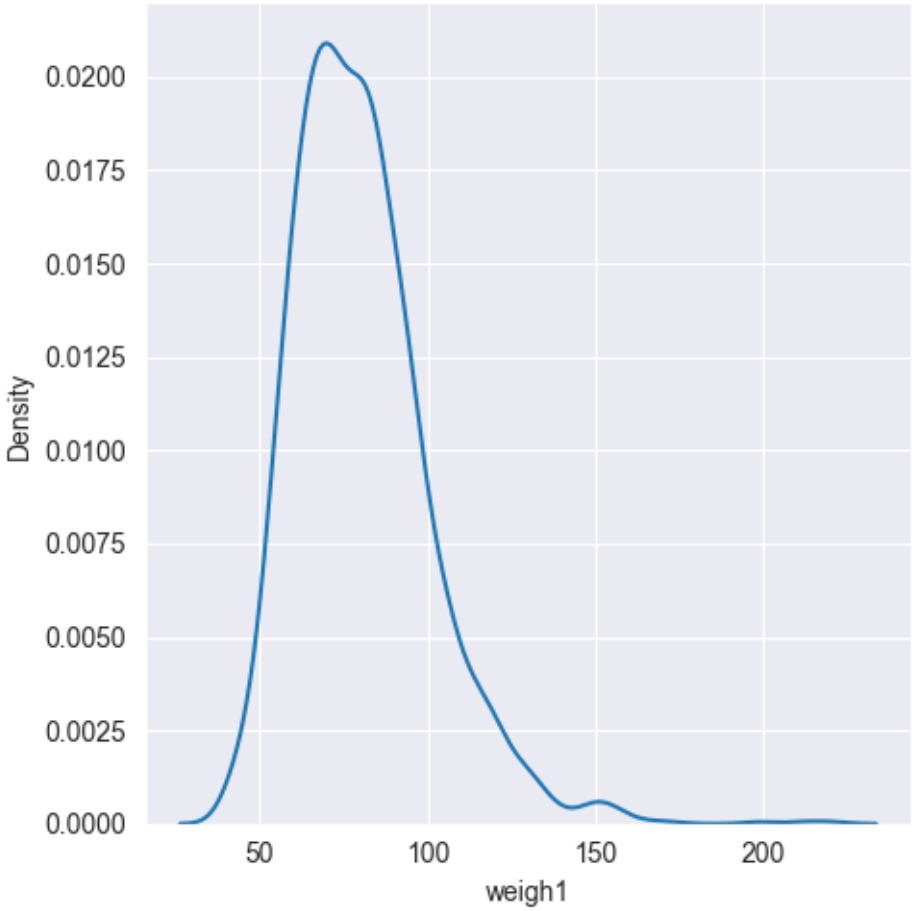
[36 rows x 89 columns]

1.4 Height distribution



geschlecht	count	mean	min	25%	50%	75%	max	std
0.0	431.0	167.765661	149.0	163.0	168.0	172.0	220.0	6.876541
1.0	244.0	179.061475	120.0	174.0	179.5	185.0	204.0	9.006189
2.0	3.0	160.000000	120.0	145.0	170.0	180.0	190.0	36.055513

1.5 Weight distribution



geschlecht	count	mean	min	25%	50%	75%	max	std
0.0	429.0	73.125874	40.0	61.0	68.0	82.0	172.0	18.461959
1.0	244.0	89.229508	50.0	75.0	87.0	98.0	154.0	18.174412
2.0	3.0	91.333333	42.0	54.5	67.0	116.0	165.0	65.010256

1.6 Sensor and app data

```
n Geographical data in baseline:          1366

tracking permission ratio: 0.756786703601108
-------------------------------------------------
n Geographical data in followup:          4673

tracking permission ratio:        0.7927056827820187

n permissions apps tracking in baseline:          101
```

```
tracking permission ratio 0.05595567867036011
------------------------------------------------
n permissions apps tracking in followup:         370

tracking permission ratio:       0.06276505513146735
```

1.7 Hypertension, Diabetes, Blood Pressure

Bei Hypertension, Diabetes und Blood Pressure interessiert uns, wie viele Patienten hiervon betroffen sind, nicht nur, wie viele diese Frage insgesamt ausgefüllt haben. Demnach bräuchten wir die Anzahl N für Patienten mit: hyper1 == 1 diabe1 == 1 blood1 == 1 Kannst du diese Abfrage noch machen?

1.7.1 Hypertension

```
Absolut numbers
 hyper1
0     1117
1      454
99     234
Name: count, dtype: int64
Relative numbers
 hyper1
0     0.618837
1     0.251524
99    0.129640
Name: proportion, dtype: float64
```

1.7.2 Diabetes

```
Absolut numbers
 diabe1
0     1453
99     239
1      113
Name: count, dtype: int64
Relative numbers
 diabe1
0     0.804986
99    0.132410
1     0.062604
Name: proportion, dtype: float64
```

1.7.3 Bloodpressure

```
Absolut numbers
 blood1
0     1071
```

```
1      461
99     273
Name: count, dtype: int64
Relative numbers
 blood1
0     0.593352
1     0.255402
99    0.151247
Name: proportion, dtype: float64
```

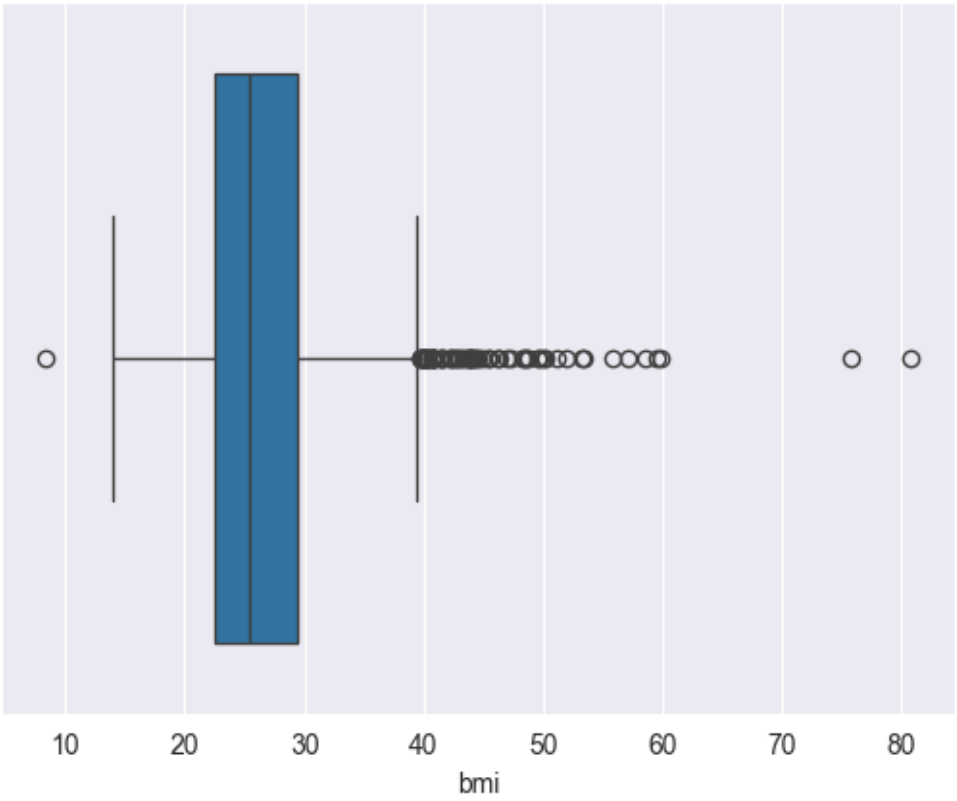
1.8 BMI

BMI haben wir von: 429 + 244 = 673 D.h. diese Anzahl an Personen hat sowohl Height als auch Weight angegeben? Aktuell ist das als zwei Grafen dargestellt, wichtig wäre zu wissen, wer beide Angaben abgegeben hat.

```
<class 'pandas.core.frame.DataFrame'>
Index: 1772 entries, 30879 to 30576
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   weigh1  1772 non-null   float64
 1   heigh   1772 non-null   float64
dtypes: float64(2)
memory usage: 41.5 KB
```

1772 Personen haben sowohl Gewicht als auch Körpergröße angegeben

```
count    1772.000000
mean       26.658209
std         6.232164
min         8.264463
25%        22.515191
50%        25.432398
75%        29.377583
max        80.808081
Name: bmi, dtype: float64
```



Und auf welche Grundgesamtheit, also Anzahl Personen im Datensatz, bezieht sich das insgesamt? Und wenn ich es richtig sehe, haben wir die Angabe egtl. von mehr Personen, aber da Geschlecht bis Dez 2020 fehlt, kann ich die Verteilung hier nicht zuordnen, korrekt? Die Werte sind ja nur aussagekräftig, wenn ich die Geschlechtsangaben habe.

Ich weiß nicht ganz, was Du meinst, aber ich hoffe, dasss die obigen Zahlen die Antwort liefen. Zu 1772 haben wir Gewicht und Körpergröße, zu 684 Personen haben wir Age, Sex, und Country.

1.9 Age, Sex, Country

Die Angaben für Alter, Geschlecht und Land haben wir erst ab 12-2020. Das sollten wir auf jeden Fall jetzt in alle weiteren Diskussionen miteinbeziehen. Entweder man fokussiert sich auf ein Thema, wo diese Faktoren eine untergeordnete Rolle spielen, oder man kann erst die Daten ab Dez 2020 verwenden. Wie viele Patienten haben wir denn für den Baselinebogen, wenn wir erst ab Dez 2020 rechnen?

Korrekt, ja. Von den 1805 Nutzern haben 1121 keine Alters-, Geschlecht- oder Länderangabe. Alter, Geschlecht und Land ist bekannt von 684 Personen.