

МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра Информационных систем

ОТЧЕТ

по практической работе
по дисциплине «Машинное обучение»

**Тема: «Разбор и анализ датасета, являющегося результатом
исследования сна человека с помощью различных устройств.»**

Студент гр. 2373

Панина А. Л.

Преподаватель

Татчина Я. А.

Санкт-Петербург

2024

Часть 1. Разбор первичного датасета

Начальный датафрейм. Вся информация хранится в колонке ‘data’, она хранится в json формате, который нужно распаковать.

0	applehealth_sleep_2261472_58c9063982329c8111fd...																	applehealth	2261472	
1	applehealth_sleep_2261472_a547bfb3709c8af9b90b...																	applehealth	2261472	
2	applehealth_sleep_2261472_e3ec8989eb733107ab6b...																	applehealth	2261472	

Получается датафрейм:

0	category	sleep	ABCD13F9-CAB2-42E8-BEFB-63C855FC3307	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...			17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	HealthKit	...				NaN	NaN		
1	category	sleep	4CDBEAB5-8229-452C-831E-3436152F94CA		NaN	com.tantsissa.AutoSleep		18.0.0	iPhone15,2	AutoSleep	6.11.21	HealthKit	...				870.0	2670.0		
2	category	sleep	BAE77E9A-1490-4901-B42D-E6026A80E97B		NaN	com.tantsissa.AutoSleep		18.0.0	iPhone15,2	AutoSleep	6.10.30	HealthKit	...				5340.0	1500.0		
3	category	sleep	BCCD626F-CD18-42A7-BB85-D1675EB1CE74		NaN	com.tantsissa.AutoSleep		17.5.1	iPhone15,2	AutoSleep	6.10.30	HealthKit	...				2010.0	2250.0		
4	category	sleep	101079AB-EC9E-483C-9429-6762D23A4271	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...			17.3.0	iPhone15,2	iPhone Аннушка	17.3	HealthKit	...				NaN	NaN		

В датасете сразу видно, что много пропущенных значений.

Колонки:

```
Index(['dataType', 'entryType', 'health_kit_id', 'HKTimeZone',
       'bundleIdentifier', 'operatingSystemVersion', 'productType',
       'sourceName', 'version', 'sourceGroup', 'sourceType', 'timeEnd',
       'timeStart', 'value', 'Asleep', 'Average HR', 'Average RespRate',
       'Average SpO2', 'Daytime HR', 'Deep Sleep', 'Energy Threshold',
       'Lights', 'Max RespRate', 'Max SpO2', 'Min RespRate', 'Min SpO2',
       'Rating', 'Recharge', 'stagesAwake', 'stagesDeep', 'stagesLight',
       'stagesREM', 'stagesSleep', 'usingStages', 'HKExternalUUID',
       'HKMetadataKeySyncIdentifier', 'HKMetadataKeySyncVersion',
       'HKWasUserEntered'],
      dtype='object')
```

Пропущенные

значения:

	0
dataType	0
entryType	0
health_kit_id	0
HKTimeZone	92
bundleIdentifier	0
operatingSystemVersion	0
productType	0
sourceName	0
version	0
sourceGroup	0
sourceType	0
timeEnd	0
timeStart	0
value	0
Asleep	154
Average HR	154
Average RespRate	154
Average SpO2	207
Daytime HR	154
Deep Sleep	154

Energy Threshold	154
Lights	154
Max RespRate	154
Max SpO2	207
Min RespRate	154
Min SpO2	207
Rating	154
Recharge	154
stagesAwake	161
stagesDeep	161
stagesLight	161
stagesREM	161
stagesSleep	161
usingStages	231
HKExternalUUID	235
HKMetadataKeySyncIdentifier	235
HKMetadataKeySyncVersion	235
HKWasUserEntered	239

dtype: int64

Количество строк: 240

Пропущенных значений достаточно много. У большей части атрибутов их больше, чем в 50% записей.

Далее рассмотрим уникальность данных:

	0
dataType	1
entryType	1
health_kit_id	230
HKTimeZone	8
bundleIdentifier	7
operatingSystemVersion	13
productType	3
sourceName	7
version	20
sourceGroup	1
sourceType	1
timeEnd	228
timeStart	228
value	4
Asleep	66
Average HR	78
Average RespRate	81
Average SpO2	28
Daytime HR	80
Deep Sleep	80
Energy Threshold	1
Lights	1
Max RespRate	17
Max SpO2	4

Min RespRate	12
Min SpO2	8
Rating	80
Recharge	30
stagesAwake	55
stagesDeep	50
stagesLight	66
stagesREM	60
stagesSleep	66
usingStages	1
HKExternalUUID	5
HKMetadataKeySyncIdentifier	5
HKMetadataKeySyncVersion	5
HKWasUserEntered	1

dtype: int64

В датасете есть колонки с одинаковыми значениями: dataType, entryType, sourceGroup, sourceType, Energy Threshold, Lights, usingStages, HKWasUserEntered. Такие признаки не содержат информации, поэтому удалим их.

	health_kit_id	HKTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value	...	Rating	Recharge	stage
0	ABCD13F9-CAB2-42E8-BEFB-63C855FC3307	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	2024-05-07T08:25:29+0200	2024-05-07T00:22:00+0200	in_bed	...	NaN	NaN	
1	4CDBEAB5-8229-462C-831E-3436152F94CA	Nan	com.tantsissa.AutoSleep	18.0.0	iPhone15,2	AutoSleep	6.11.21	2024-10-31T07:00:00+0100	2024-10-30T23:53:00+0100	in_bed	...	70.73	87	
2	BAE77E9A-1490-4001-B42D-E6026AB0E97B	Nan	com.tantsissa.AutoSleep	18.0.0	iPhone15,2	AutoSleep	6.10.30	2024-10-02T06:42:00+0200	2024-10-02T00:07:00+0200	in_bed	...	54.86	74	
3	BCCD626F-CD18-42A7-8BB5-D1675EB1CE74	Nan	com.tantsissa.AutoSleep	17.5.1	iPhone15,2	AutoSleep	6.10.30	2024-06-29T08:37:00+0200	2024-06-28T23:31:00+0200	in_bed	...	75.88	99	
4	101079AB-EC9E-483C-9429-6762D23A4271	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.3.0	iPhone15,2	iPhone Аннушка	17.3	2024-02-01T05:41:17+0100	2024-01-31T22:25:00+0100	in_bed	...	NaN	NaN	

5 rows x 30 columns

	0
health_kit_id	0
HKTimeZone	92
bundleIdentifier	0
operatingSystemVersion	0
productType	0
sourceName	0
version	0
timeEnd	0
timeStart	0
value	0
Asleep	154
Average HR	154
Average RespRate	154
Average SpO2	207
Daytime HR	154
Deep Sleep	154
Max RespRate	154
Max SpO2	207
Min RespRate	154
Min SpO2	207
Rating	154
Recharge	154
stagesAwake	161
stagesDeep	161
stagesLight	161
stagesREM	161
stagesSleep	161
HKExternalUUID	235
HKMetadataKeySyncIdentifier	235
HKMetadataKeySyncVersion	235

dtype: int64

Также удалим колонки, где слишком много пропущенных значений, а именно:

HKExternalUUID, HKMetadataKeySyncIdentifier,
HKMetadataKeySyncVersion, Average SpO2, Max SpO2, Min SpO2

health_kit_id	HKTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value	...	Deep Sleep	Max RespRate	Min RespRate
0	ABCD13F9-CAB2-42E8-BEFB-63C855FC3307	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	2024-05-07T08:25:29+0200	2024-05-07T00:22:00+0200	in_bed	...	NaN	NaN
1	4CDBEAB5-8229-452C-831E-3436152F94CA	Nan	com.tantsissa.AutoSleep	18.0.0	iPhone15,2	AutoSleep	6.11.21	2024-10-31T07:00:00+0100	2024-10-30T23:53:00+0100	in_bed	...	12510	22.5000
2	BAE77E9A-1490-4901-B42D-E6026AB0E97B	Nan	com.tantsissa.AutoSleep	18.0.0	iPhone15,2	AutoSleep	6.10.30	2024-10-02T06:42:00+0200	2024-10-02T00:07:00+0200	in_bed	...	8738	23.0000
3	BCCD626F-CD18-42A7-8BB5-D1675EB1CE74	Nan	com.tantsissa.AutoSleep	17.5.1	iPhone15,2	AutoSleep	6.10.30	2024-06-29T08:37:00+0200	2024-06-28T23:31:00+0200	in_bed	...	9500	18.5000
4	101079AB-EC9E-483C-9429-6762D23A4271	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.3.0	iPhone15,2	iPhone Аннушка	17.3	2024-02-01T05:41:17+0100	2024-01-31T22:25:00+0100	in_bed	...	NaN	NaN

5 rows x 24 columns

	0
health_kit_id	0
HKTimeZone	92
bundleIdentifier	0
operatingSystemVersion	0
productType	0
sourceName	0
version	0
timeEnd	0
timeStart	0
value	0
Asleep	154
Average HR	154
Average RespRate	154
Daytime HR	154
Deep Sleep	154
Max RespRate	154
Min RespRate	154
Rating	154
Recharge	154
stagesAwake	161
stagesDeep	161
stagesLight	161
stagesREM	161
stagesSleep	161

Удалим строки с большим количеством пропущенных значений

Разбор df_1

	health_kit_id	HKTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value	...	Deep Sleep	Max RespRate	Min RespRate	R
1	4CDBEAB5-8229-452C-831E-3436152F94CA	NaN	com.tantsissa.AutoSleep	18.0.0	iPhone15,2	AutoSleep	6.11.21	2024-10-31T07:00:00+0100	2024-10-30T23:53:00+0100	in_bed	...	12510	22.5000	15.0000	
2	BAE77E9A-1490-4901-B42D-E6026AB0E97B	NaN	com.tantsissa.AutoSleep	18.0.0	iPhone15,2	AutoSleep	6.10.30	2024-10-02T06:42:00+0200	2024-10-02T00:07:00+0200	in_bed	...	8738	23.0000	17.5000	
3	BCCD626F-CD18-42A7-8BB5-D1675EB1CE74	NaN	com.tantsissa.AutoSleep	17.5.1	iPhone15,2	AutoSleep	6.10.30	2024-06-29T08:37:00+0200	2024-06-28T23:31:00+0200	in_bed	...	9500	18.5000	15.0000	
6	D8E7F33D-3292-495E-843D-67CC4DB59C63	NaN	com.tantsissa.AutoSleep	17.5.1	iPhone15,2	AutoSleep	6.10.30	2024-07-07T07:43:00+0200	2024-07-06T23:19:00+0200	in_bed	...	9220	19.5000	14.0000	
15	355FDA47-396C-4FCB-8739-D1FE36CC56B7	NaN	com.tantsissa.AutoSleep	17.6.1	iPhone15,2	AutoSleep	6.10.30	2024-09-07T06:46:00+0200	2024-09-06T22:59:00+0200	in_bed	...	9729	20.0000	14.5000	

5 rows x 24 columns

Количество строк: 86

Пропущенные значения:

	0
health_kit_id	0
HKTimeZone	86
bundleIdentifier	0
operatingSystemVersion	0
productType	0
sourceName	0
version	0
timeEnd	0
timeStart	0
value	0
Asleep	0
Average HR	0
Average RespRate	0
Daytime HR	0
Deep Sleep	0
Max RespRate	0
Min RespRate	0
Rating	0
Recharge	0
stagesAwake	7
stagesDeep	7
stagesLight	7
stagesREM	7
stagesSleep	7

Признак HKTimeZone отсутствует у данного датасета. Удалим его.

	0
health_kit_id	85
bundleIdentifier	1
operatingSystemVersion	5
productType	1
sourceName	1
version	2
timeEnd	82
timeStart	82
value	1
Asleep	66
Average HR	78
Average RespRate	81
Daytime HR	80
Deep Sleep	80
Max RespRate	17
Min RespRate	12
Rating	80
Recharge	30
stagesAwake	55
stagesDeep	50
stagesLight	66
stagesREM	60
stagesSleep	66

dtype: int64

Удалим ненужные признаки:

- health_kit_id - так как это идентификатор;
- productType, sourceName, value - так как это уникальные значения.

Переформатируем время, так как для анализа не требуется дата, а только время отхода ко сну и время подъема. И, чтобы ранжировать время, возьмем только часы. (заодно переведем дату в числовой признак).

	operatingSystemVersion	version	timeEnd	timeStart	Asleep	Average HR	Average RespRate	Daytime Sleep	Deep Sleep	Max RespRate	Min RespRate	Rating	Recharge	stagesAwake	stagesDeep	stagesLight	stagesREM
1	18.0.0	6.11.21	7	23	25020.0	63.50	16.7564	84.27	12510	22.5000	15.0000	70.73	87	870.0	2670.0	18750.0	360
2	18.0.0	6.10.30	6	0	21360.0	73.44	19.9405	93.23	8738	23.0000	17.5000	54.86	74	5340.0	1500.0	15510.0	450
3	17.5.1	6.10.30	8	23	28500.0	66.26	16.8250	79.53	9500	18.5000	15.0000	75.88	99	2010.0	2250.0	23760.0	414
6	17.5.1	6.10.30	7	23	27660.0	68.31	16.7167	81.00	9220	19.5000	14.0000	73.31	96	1290.0	2640.0	21390.0	483
15	17.6.1	6.10.30	6	22	23880.0	68.74	16.3942	84.09	9729	20.0000	14.5000	61.70	83	2100.0	3870.0	17910.0	357

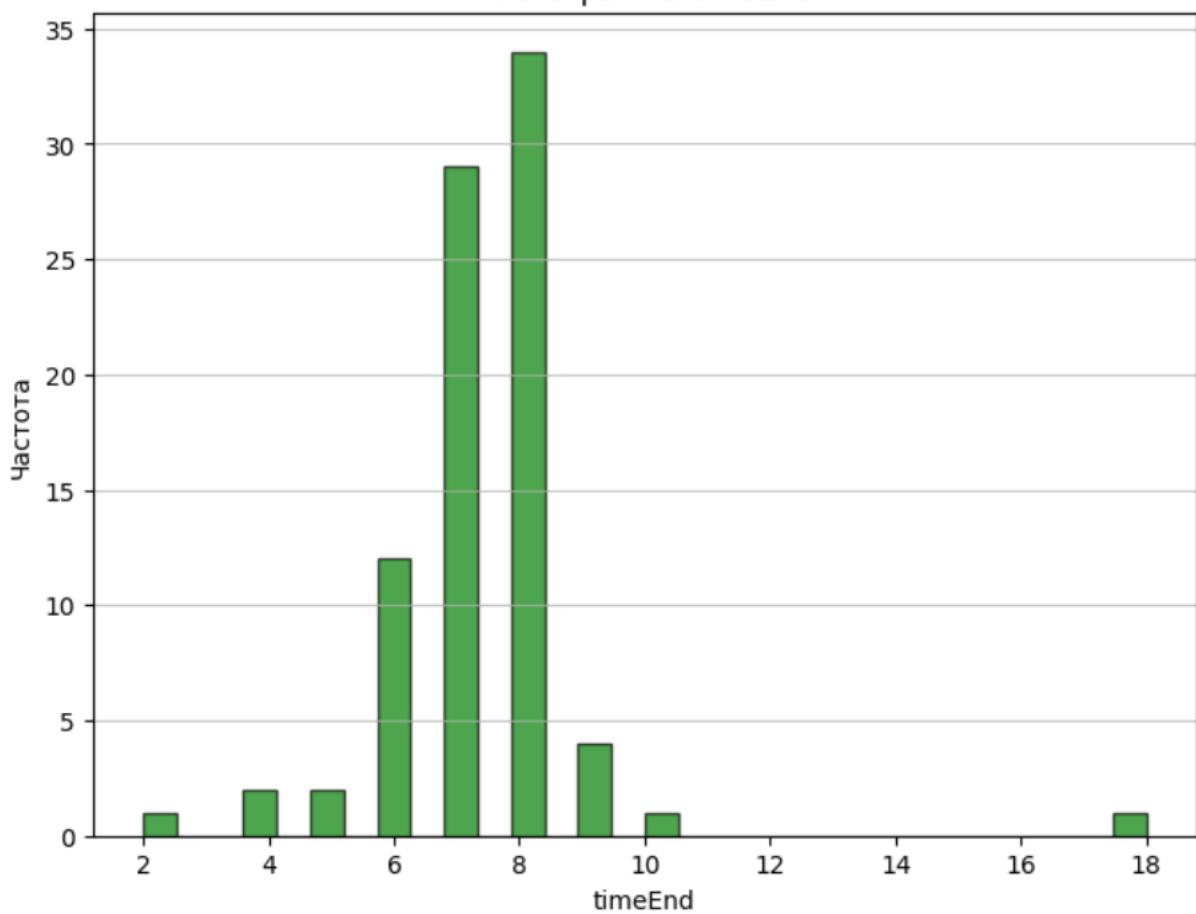
Описание числовых признаков:

	timeEnd	timeStart	Asleep	stagesAwake	stagesDeep	stagesLight	stagesREM	stagesSleep
count	86.000000	86.000000	86.000000	79.000000	79.000000	79.000000	79.000000	79.000000
mean	7.337209	13.604651	25052.093023	2234.050633	2828.734177	18694.177215	4877.468354	26400.379747
std	1.656435	11.111498	5115.182503	2094.772810	802.157934	3175.140460	1485.760160	3740.595881
min	2.000000	0.000000	9720.000000	150.000000	750.000000	9630.000000	930.000000	14250.000000
25%	7.000000	0.000000	23040.000000	900.000000	2385.000000	16620.000000	4005.000000	24075.000000
50%	7.000000	22.000000	25110.000000	1440.000000	2850.000000	18750.000000	4800.000000	27030.000000
75%	8.000000	23.000000	27555.000000	2700.000000	3435.000000	20910.000000	5910.000000	28470.000000
max	18.000000	23.000000	51420.000000	9960.000000	4290.000000	27300.000000	8520.000000	35220.000000

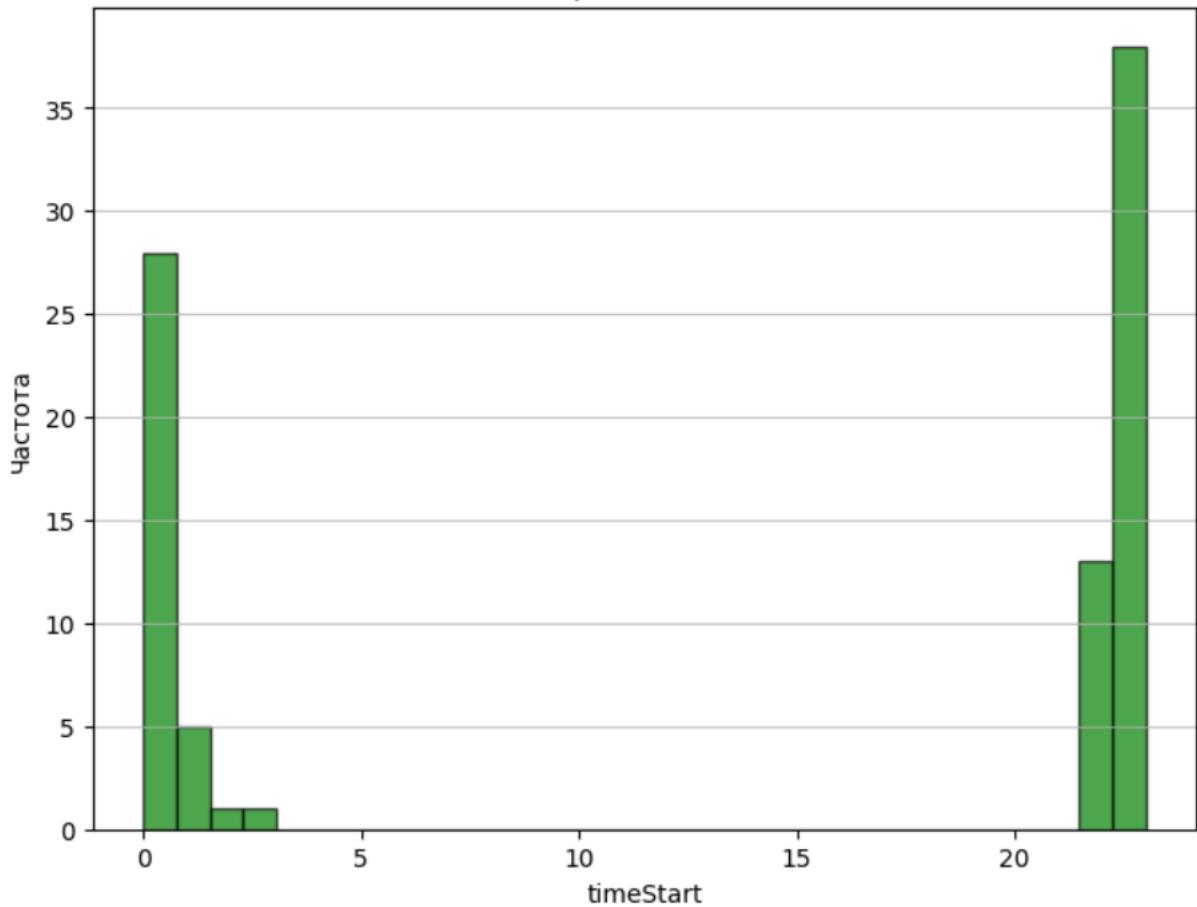
	0
operatingSystemVersion	object
version	object
timeEnd	int64
timeStart	int64
Asleep	float64
Average HR	object
Average RespRate	object
Daytime HR	object
Deep Sleep	object
Max RespRate	object
Min RespRate	object
Rating	object
Recharge	object
stagesAwake	float64
stagesDeep	float64
stagesLight	float64
stagesREM	float64
stagesSleep	float64
dtype:	object

Посмотрим на гистограммы и боксплоты.

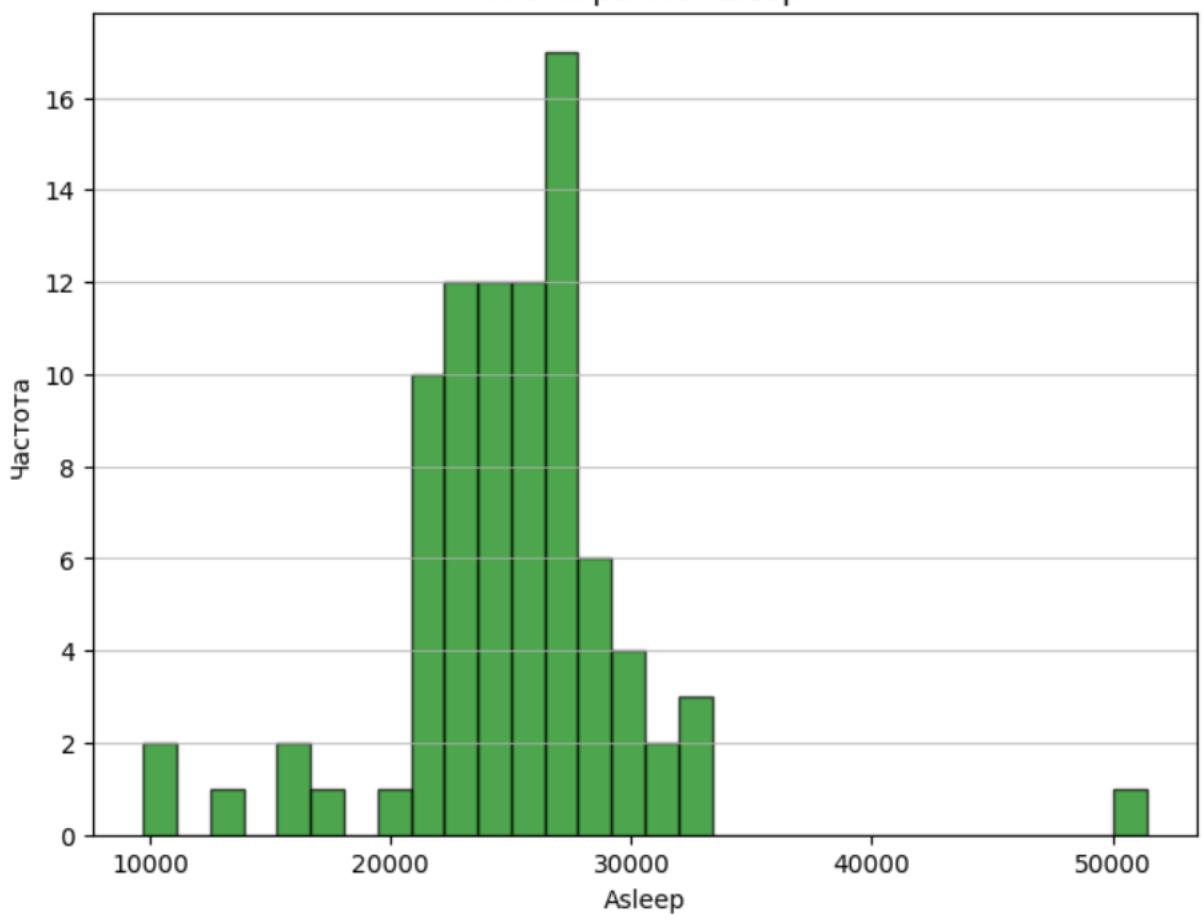
Гистограмма timeEnd



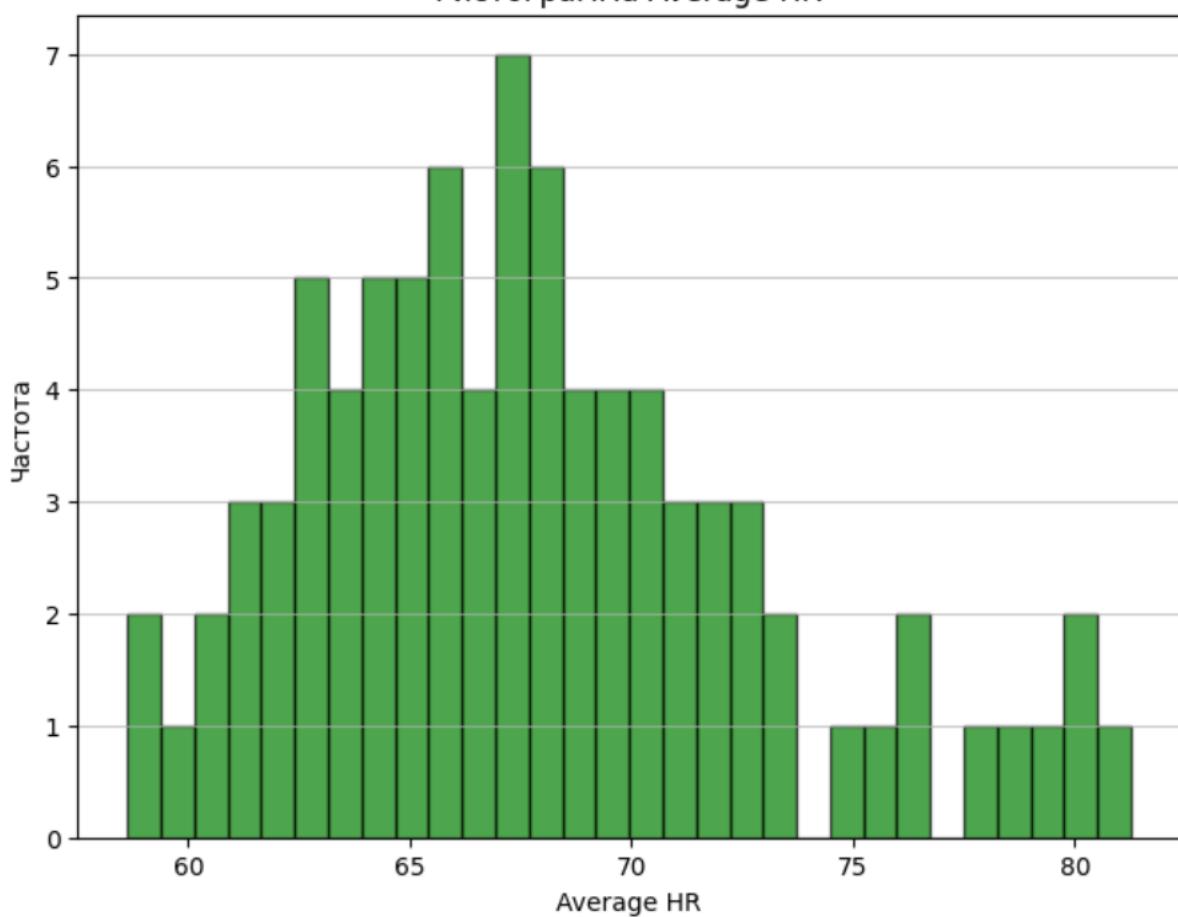
Гистограмма timeStart



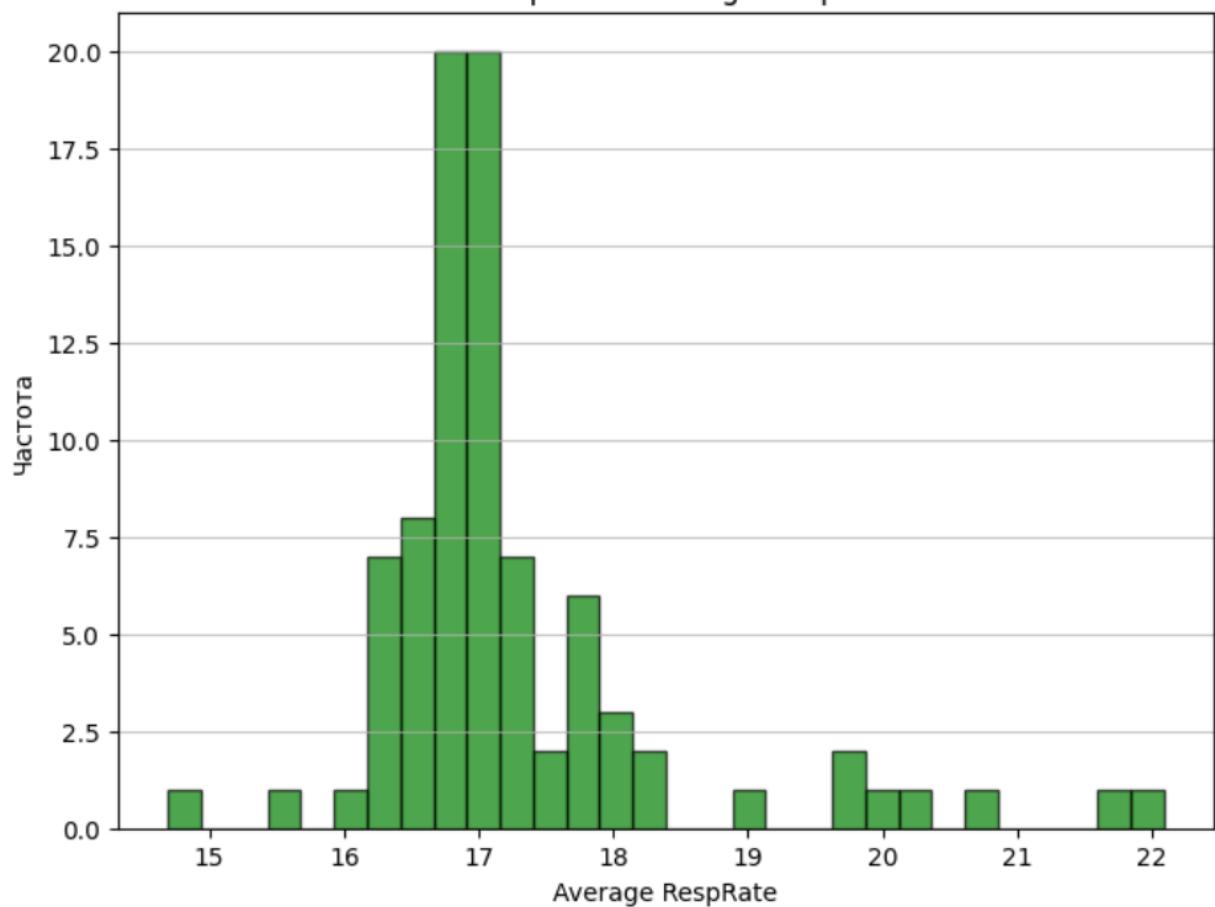
Гистограмма Asleep



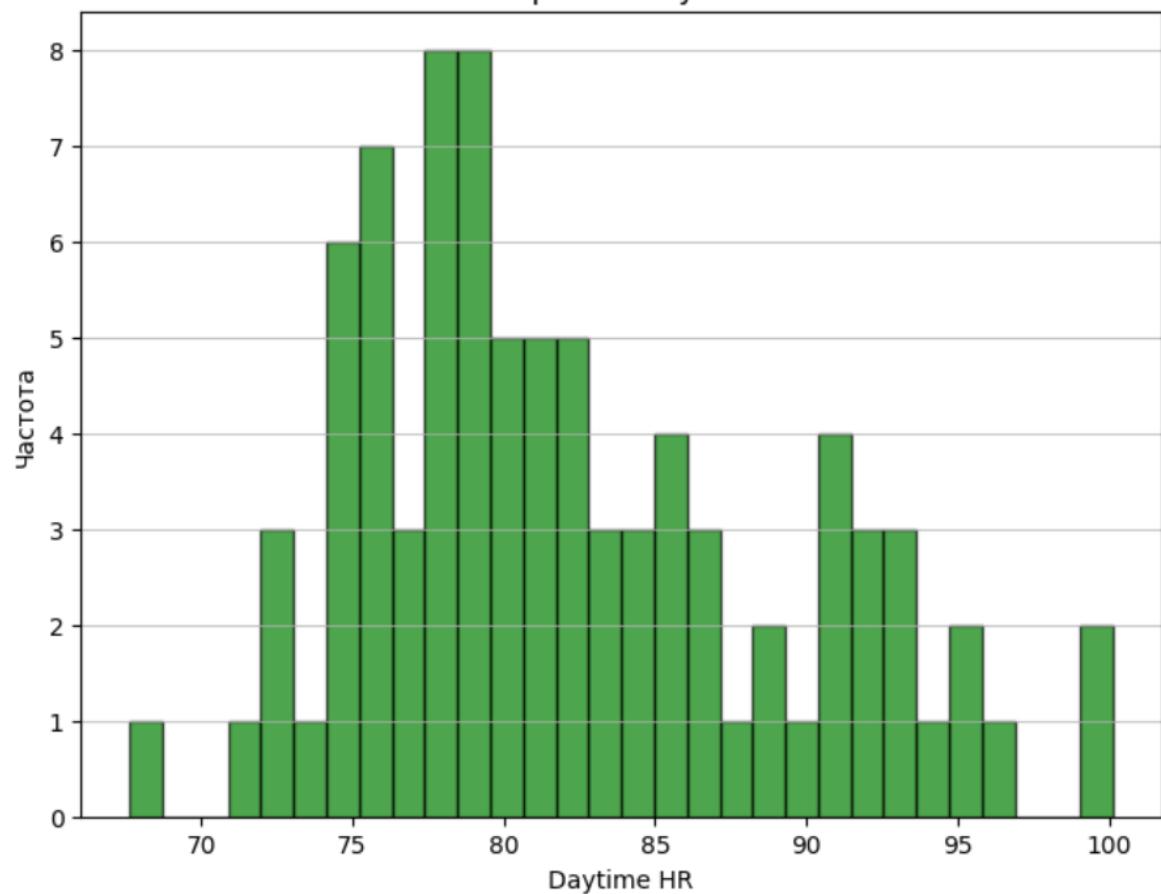
Гистограмма Average HR



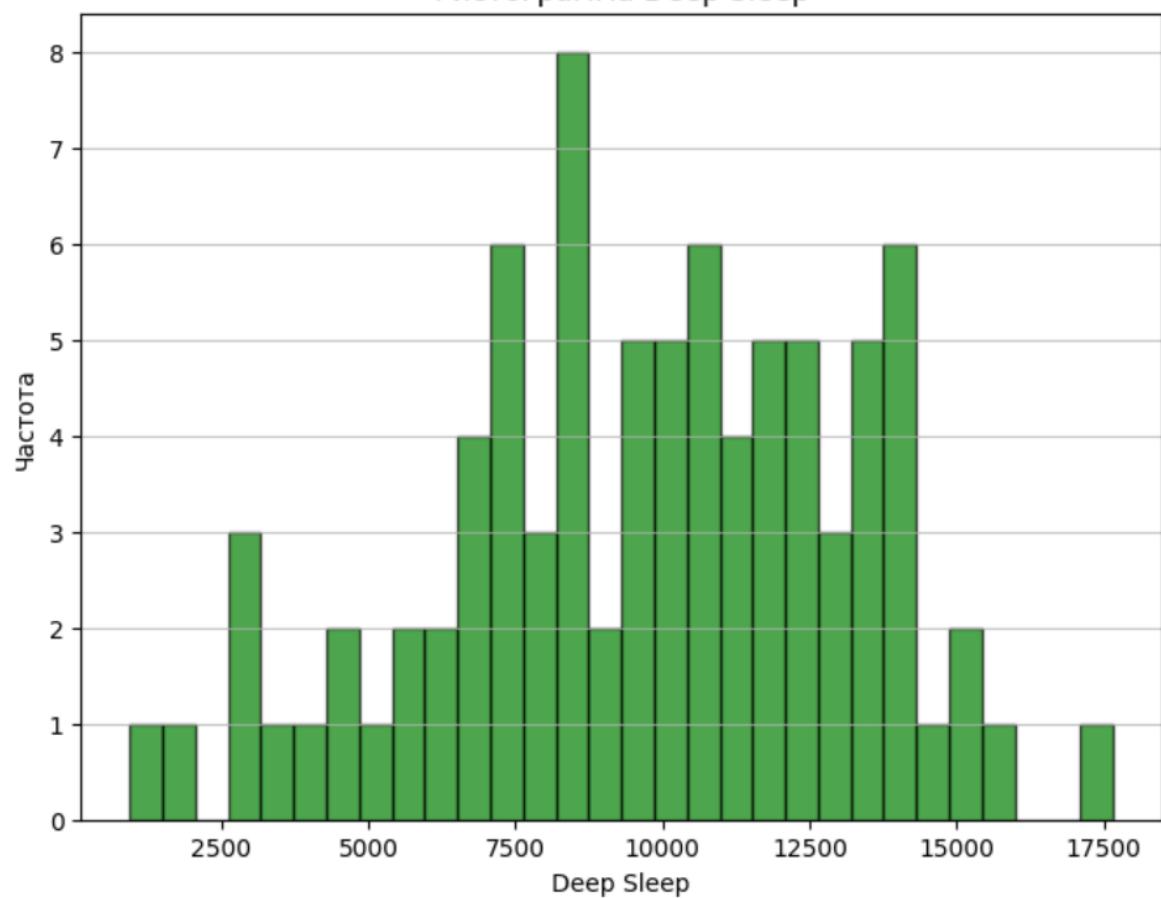
Гистограмма Average RespRate



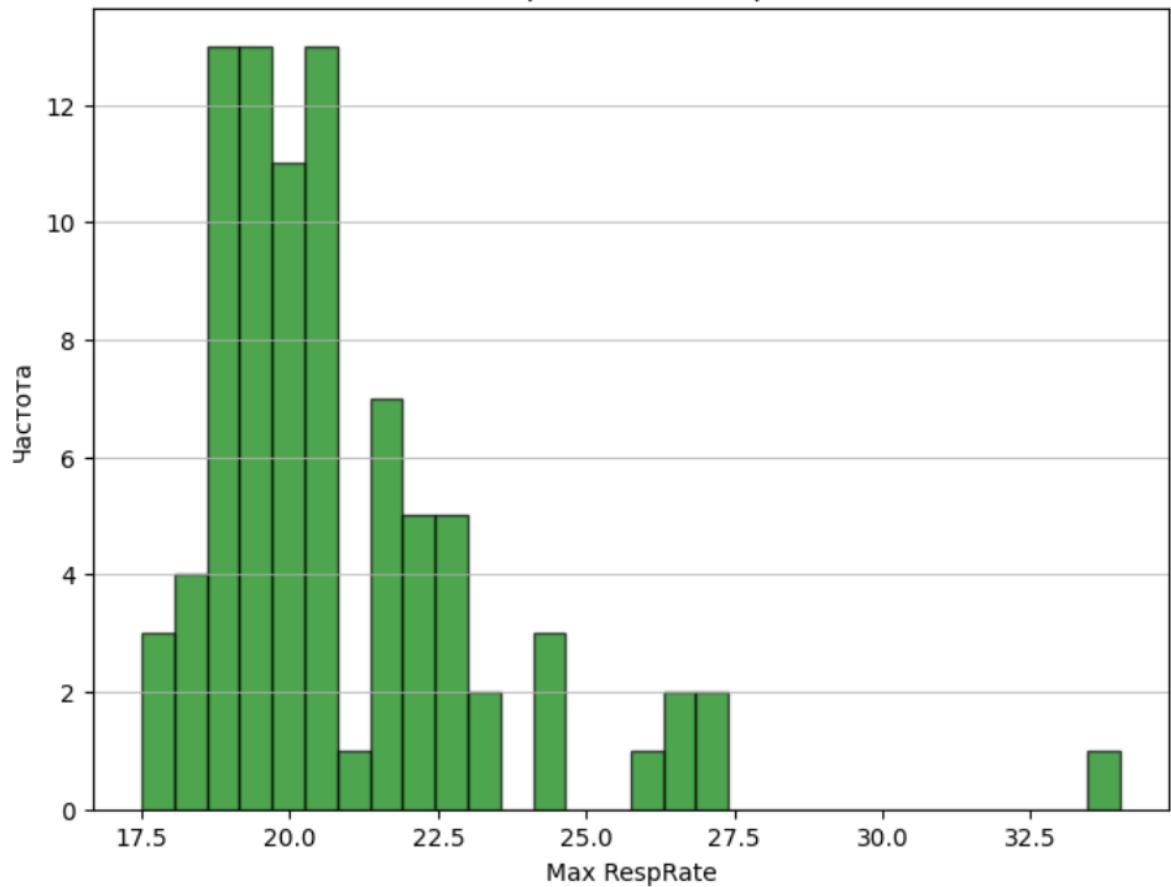
Гистограмма Daytime HR



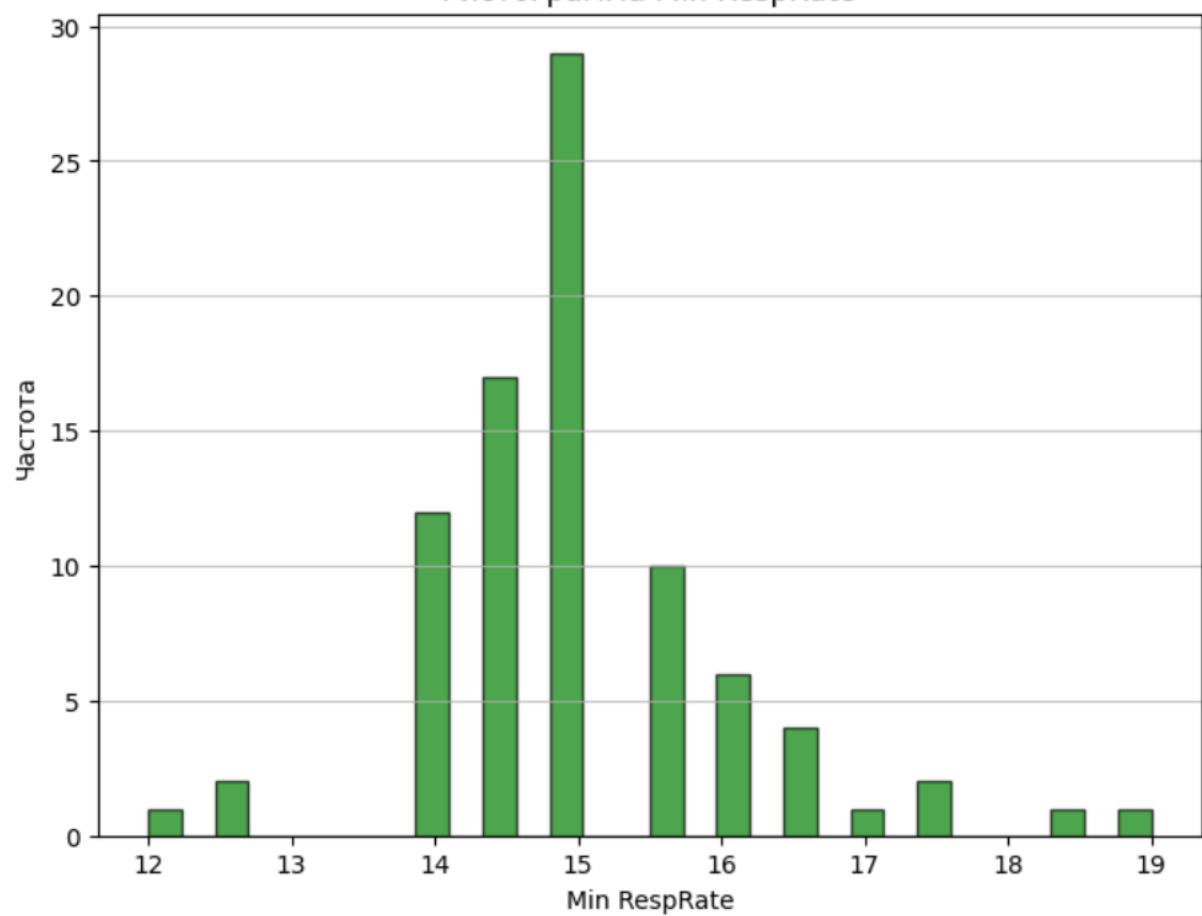
Гистограмма Deep Sleep



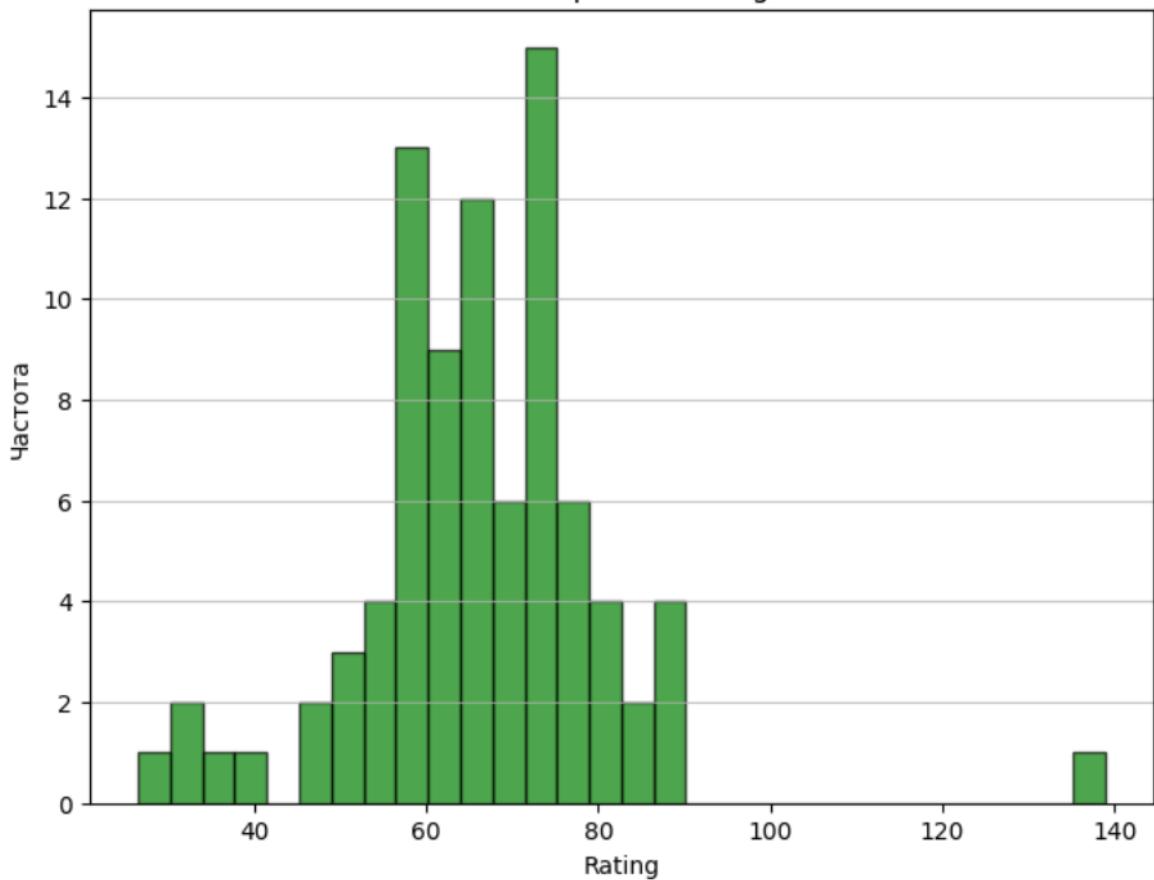
Гистограмма Max RespRate



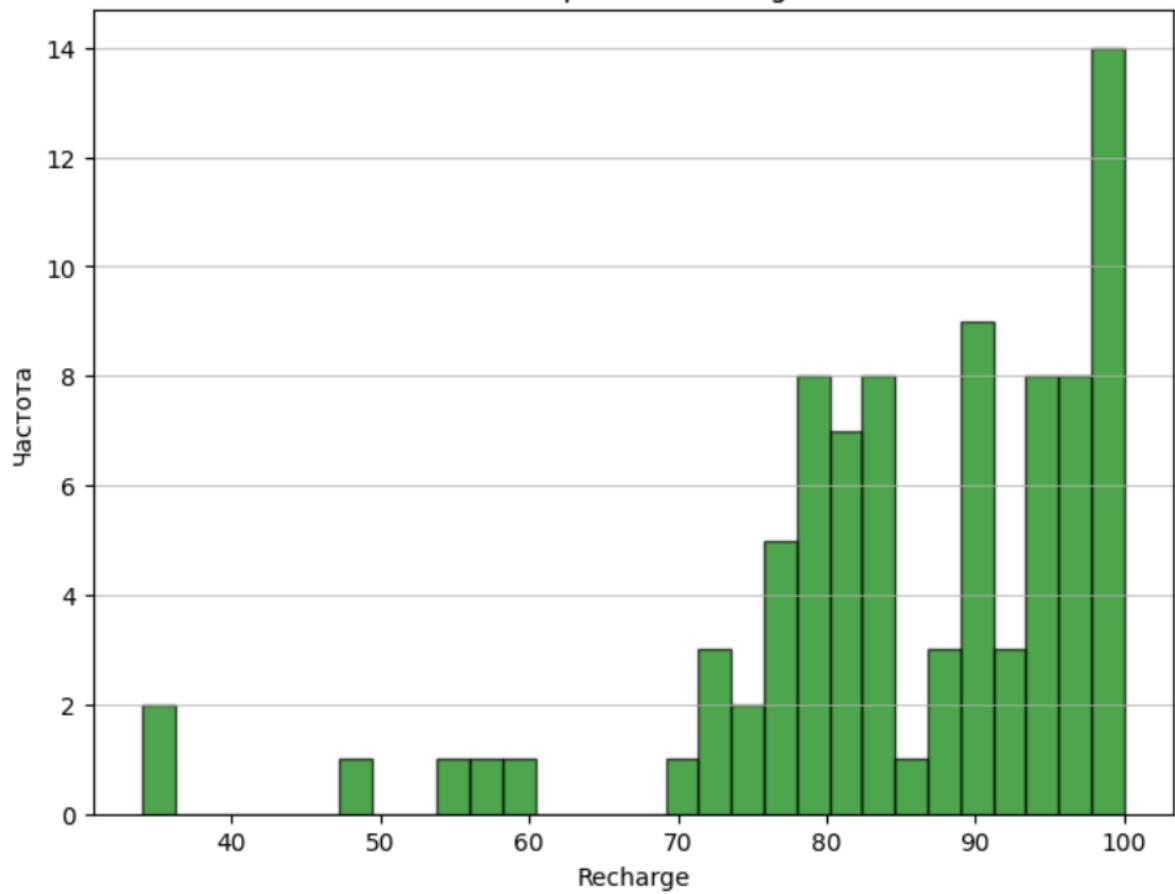
Гистограмма Min RespRate



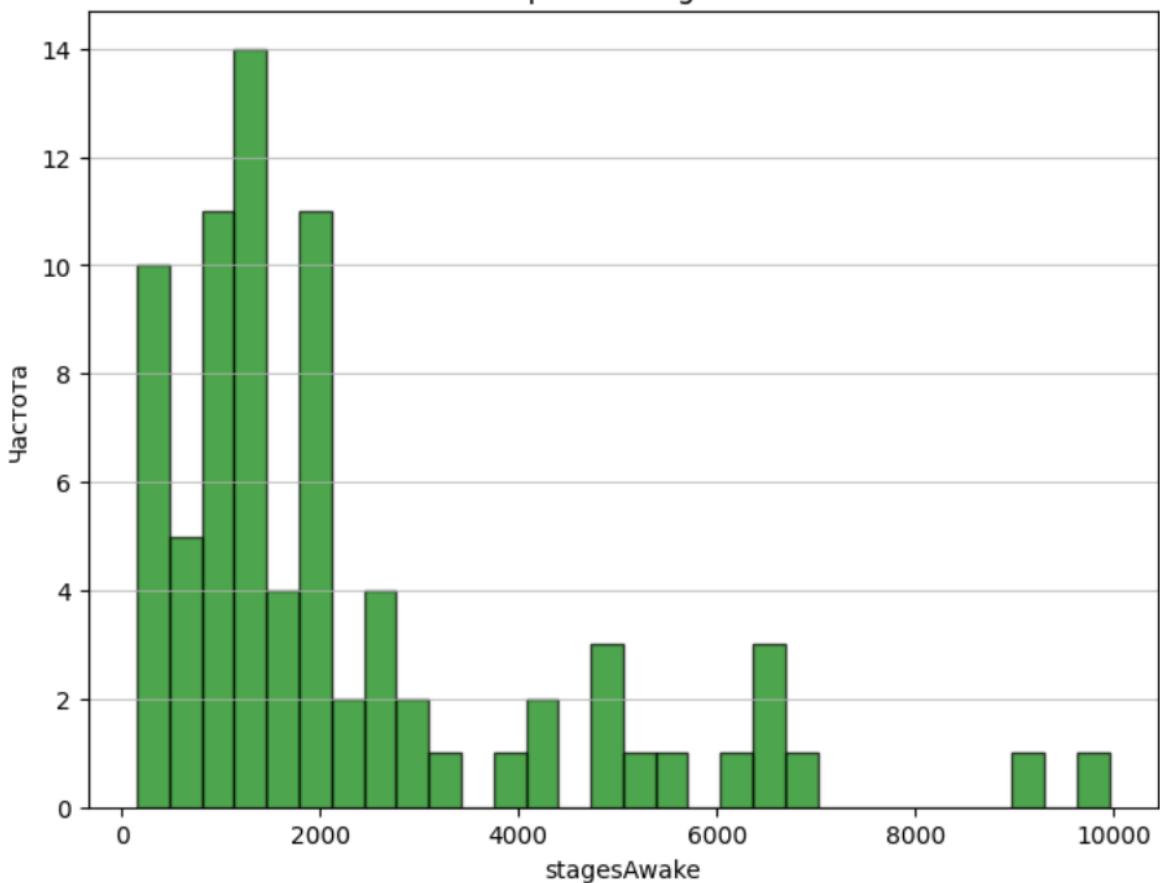
Гистограмма Rating



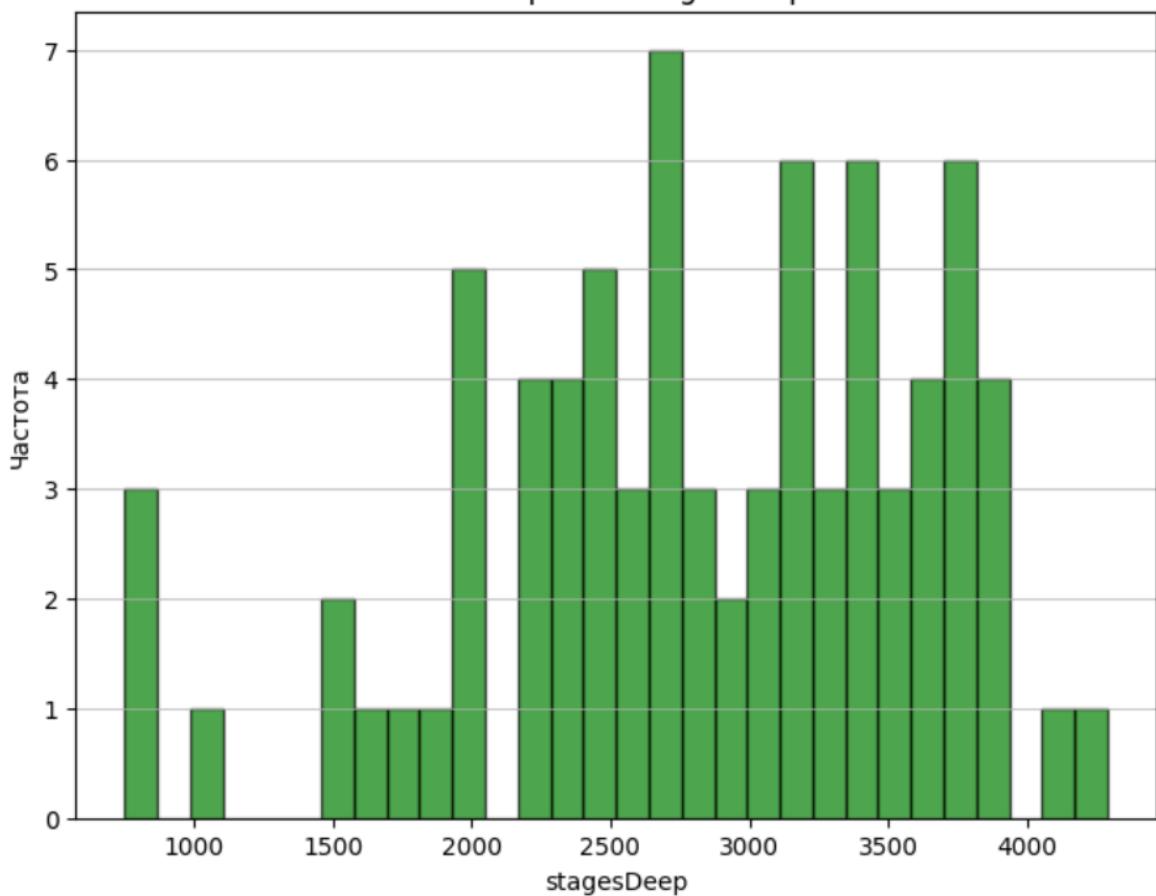
Гистограмма Recharge



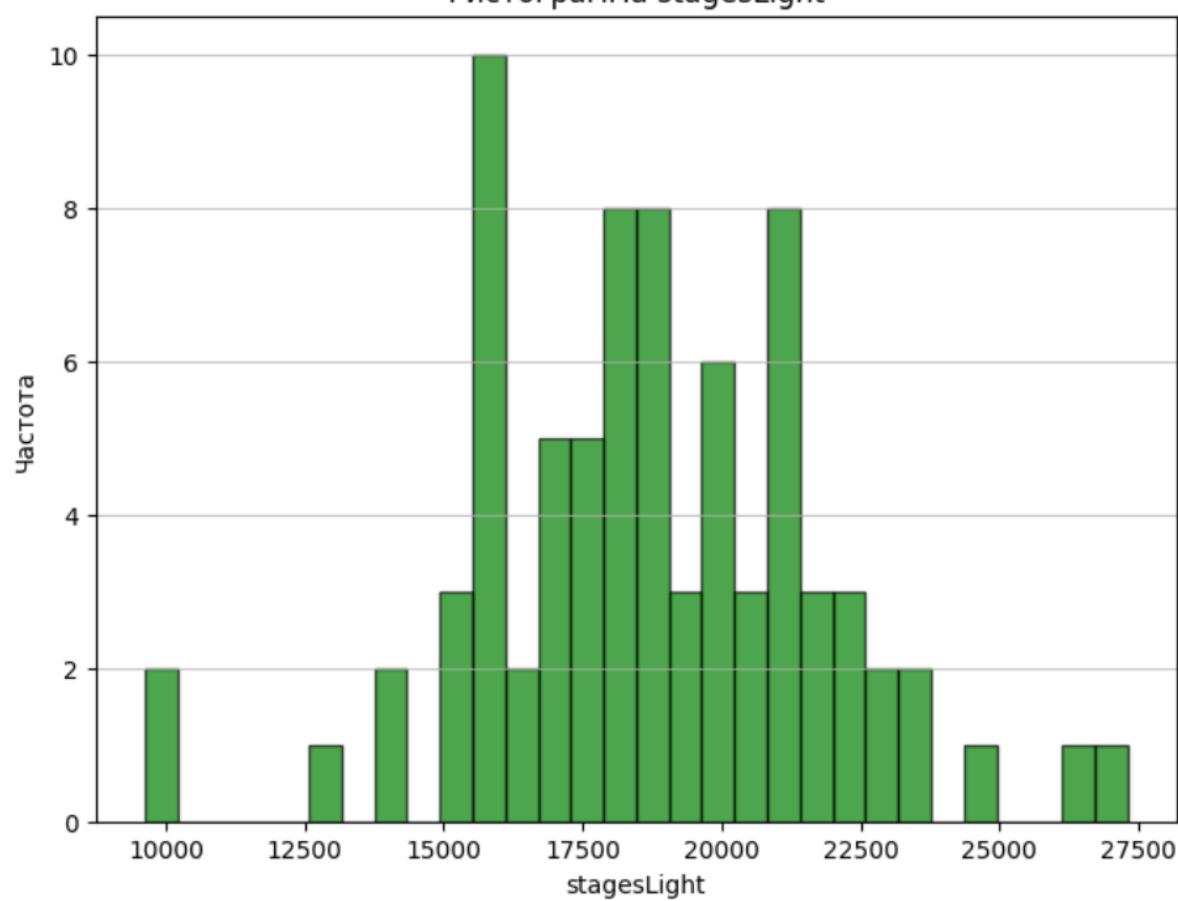
Гистограмма stagesAwake



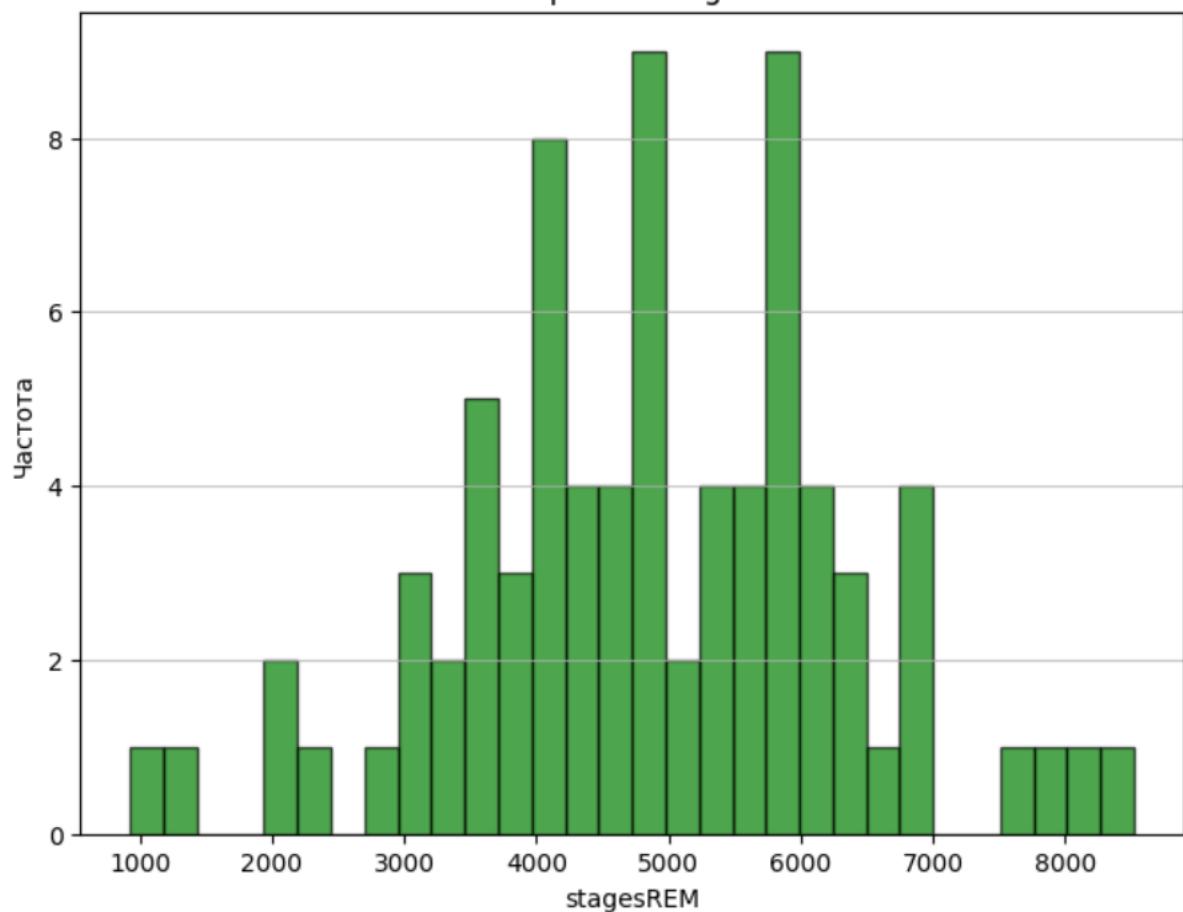
Гистограмма stagesDeep



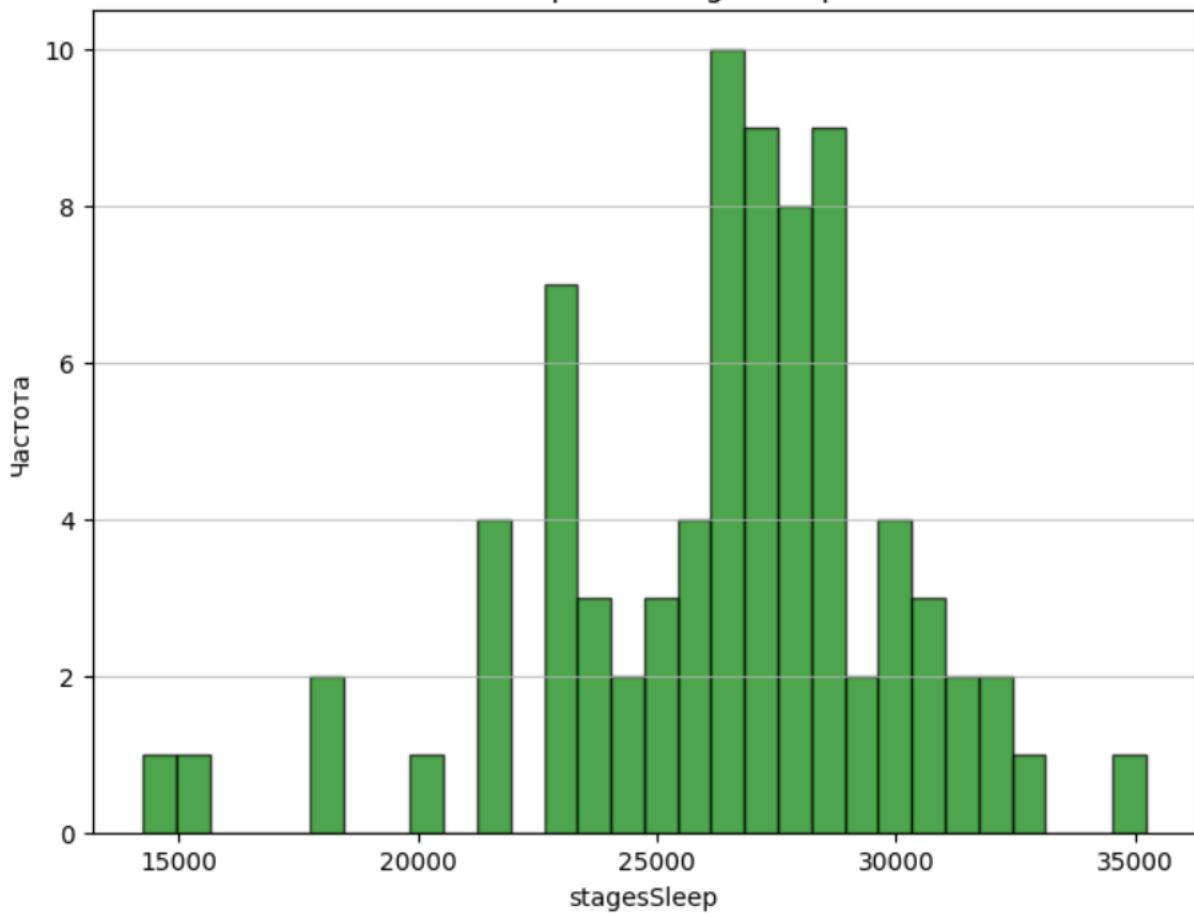
Гистограмма stagesLight



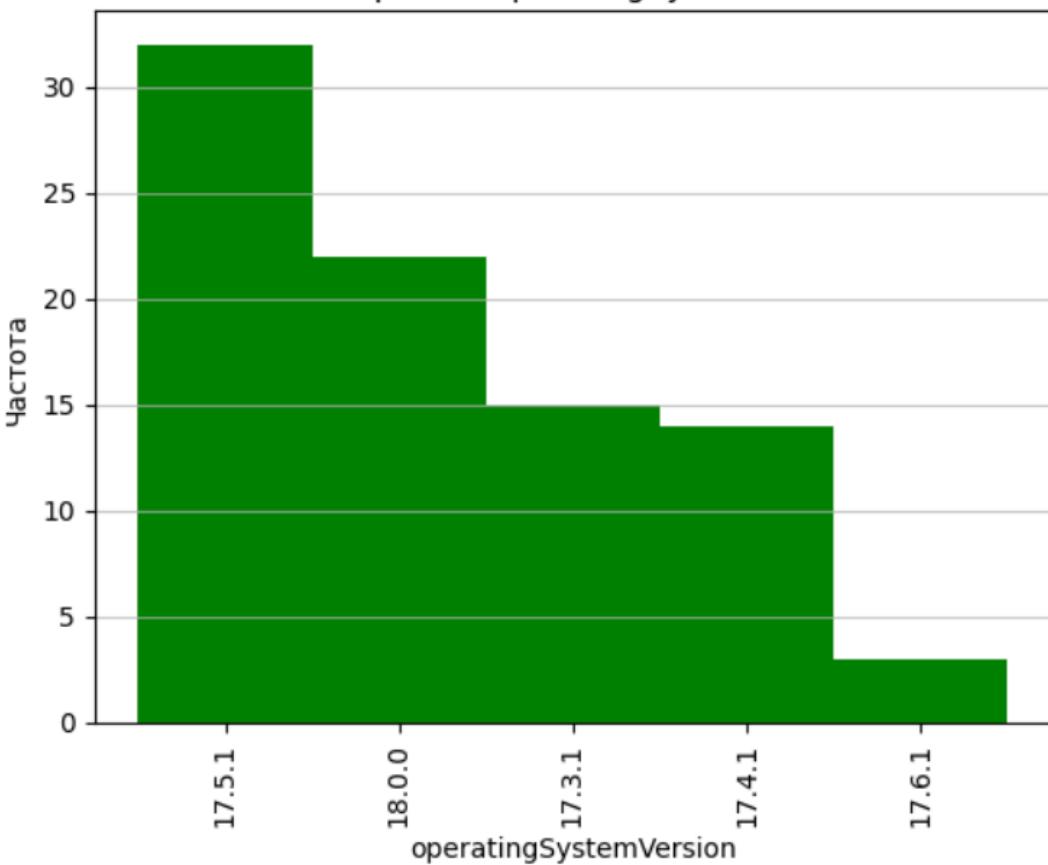
Гистограмма stagesREM



Гистограмма stagesSleep



Гистограмма operatingSystemVersion



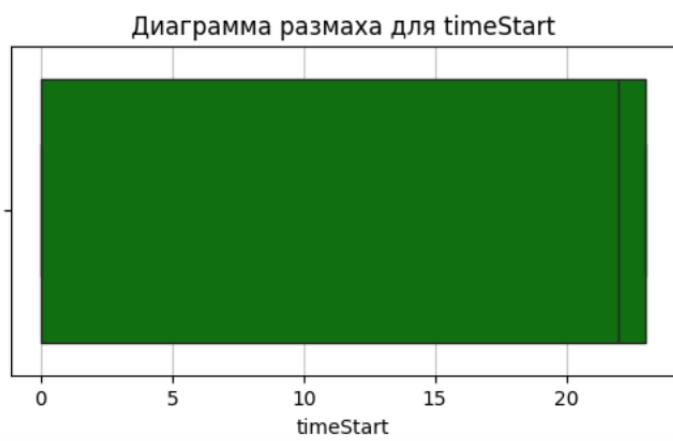
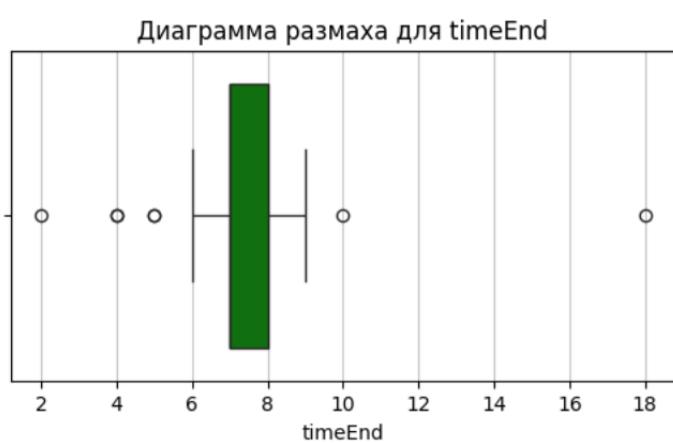
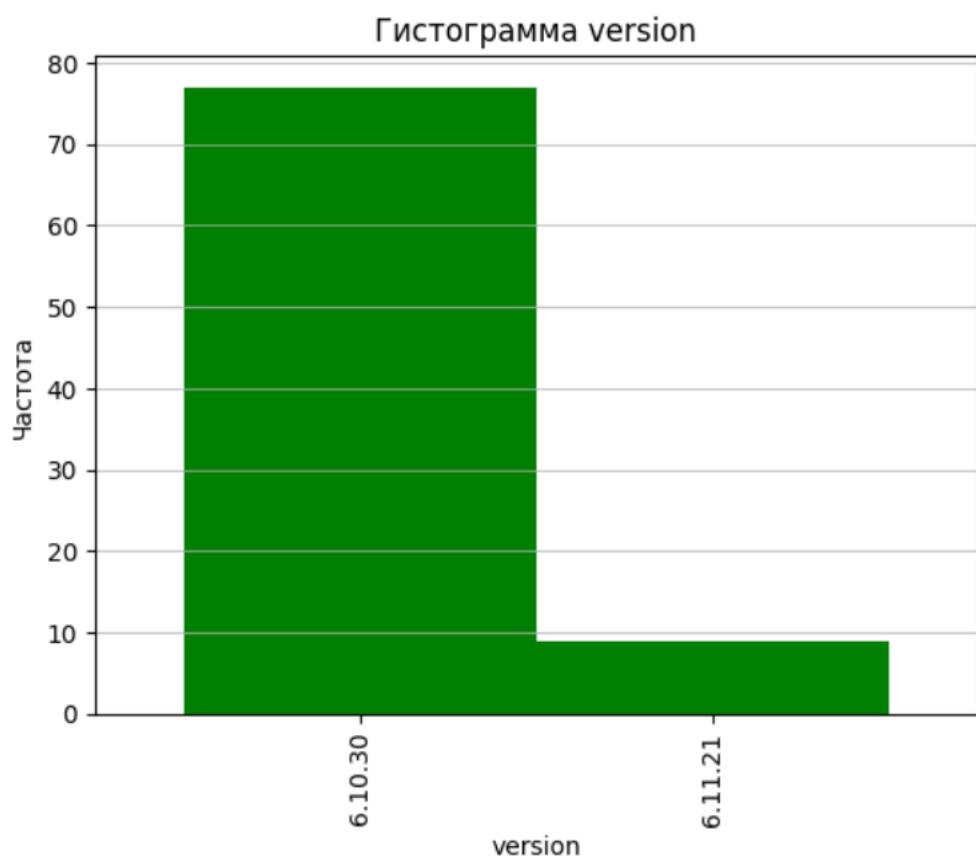


Диаграмма размаха для Asleep

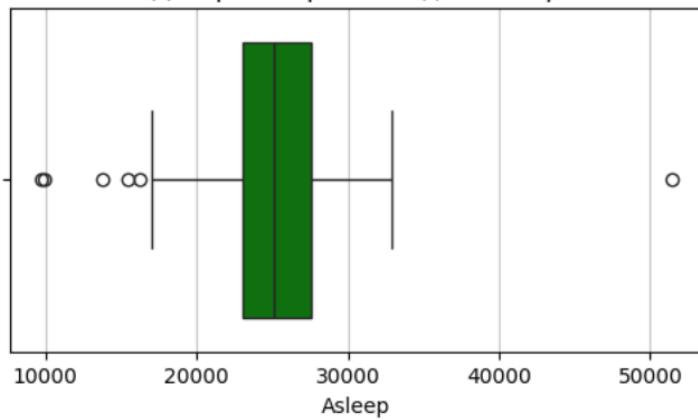


Диаграмма размаха для Average HR

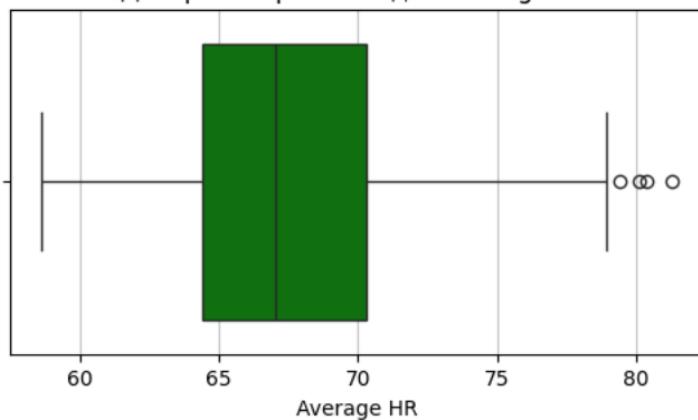


Диаграмма размаха для Average RespRate

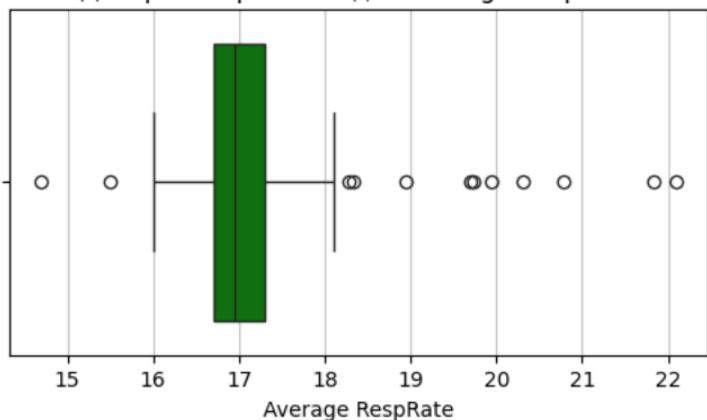


Диаграмма размаха для Daytime HR

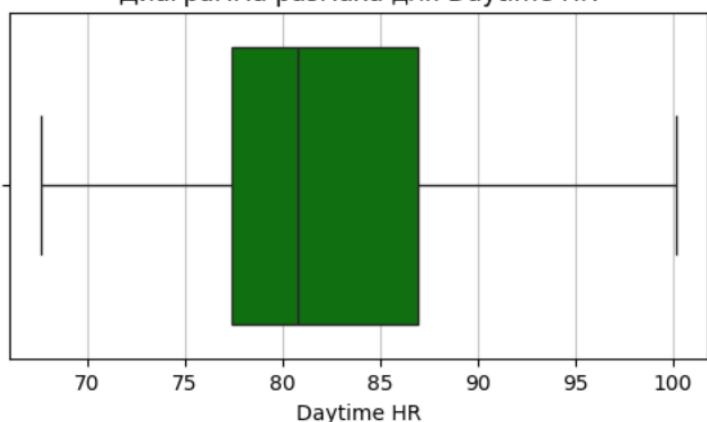


Диаграмма размаха для Deep Sleep

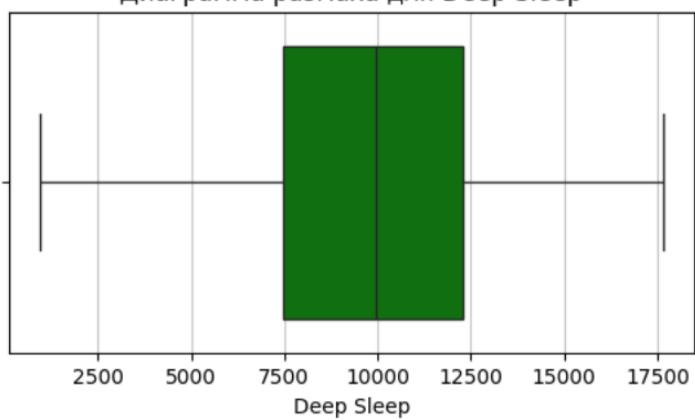


Диаграмма размаха для Max RespRate

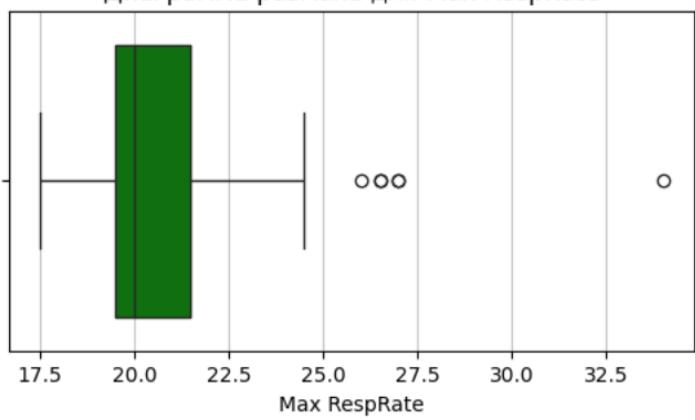


Диаграмма размаха для Min RespRate

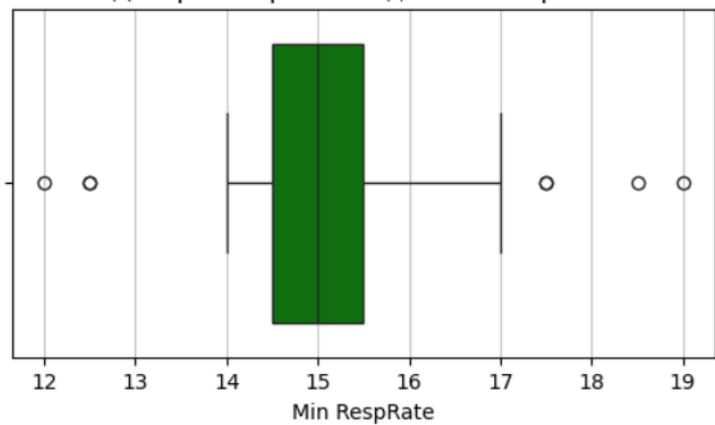


Диаграмма размаха для Rating

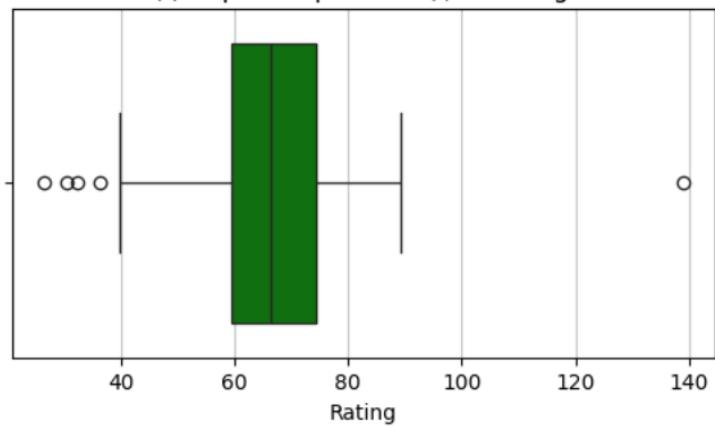


Диаграмма размаха для Recharge

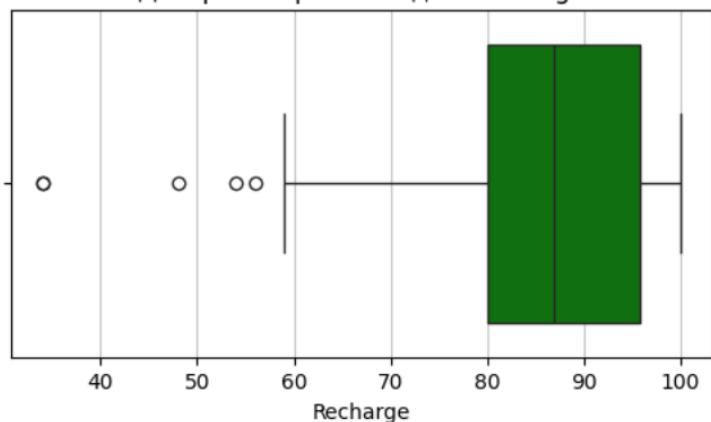


Диаграмма размаха для stagesAwake

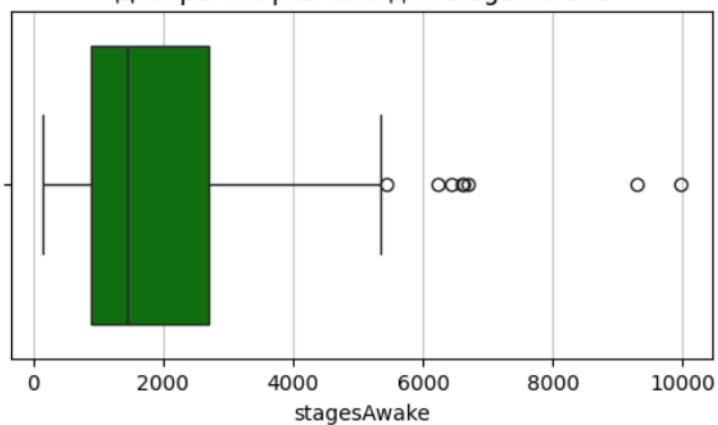


Диаграмма размаха для stagesDeep

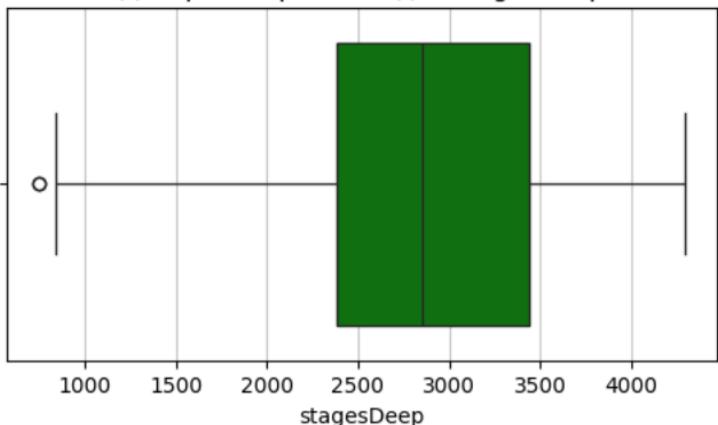


Диаграмма размаха для stagesLight

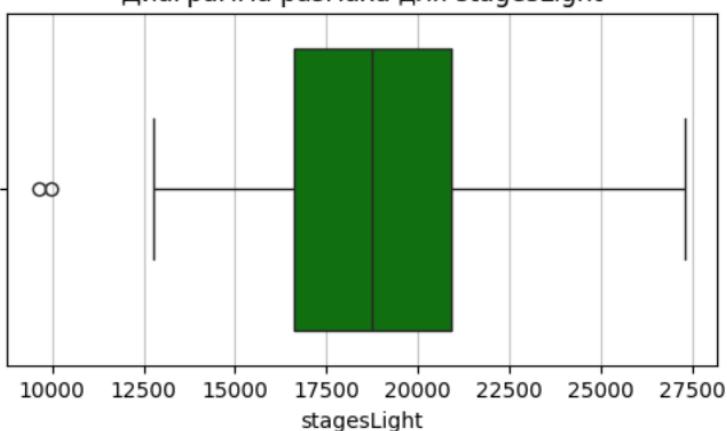


Диаграмма размаха для stagesREM

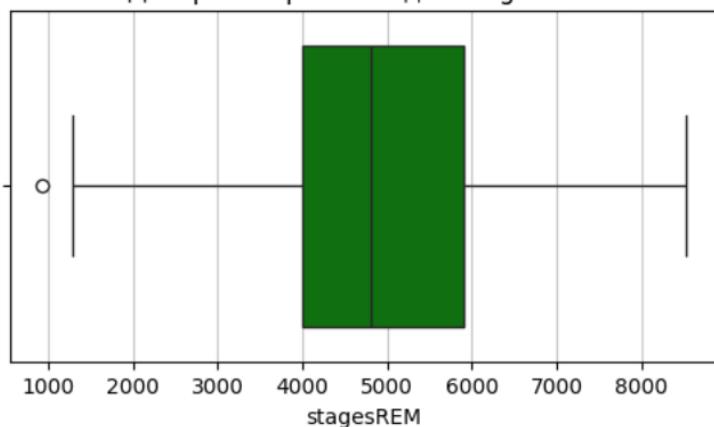
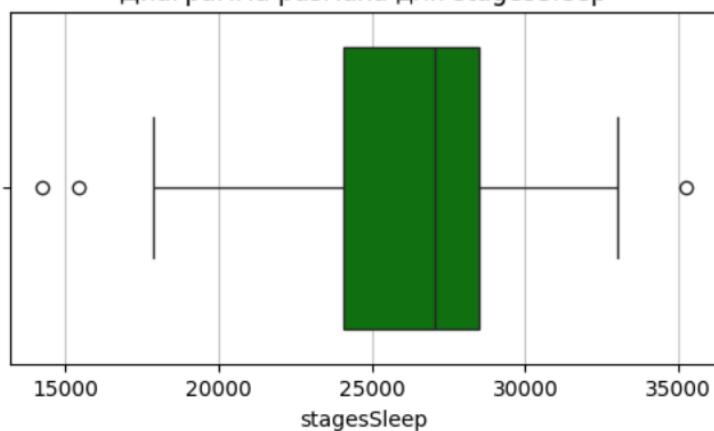


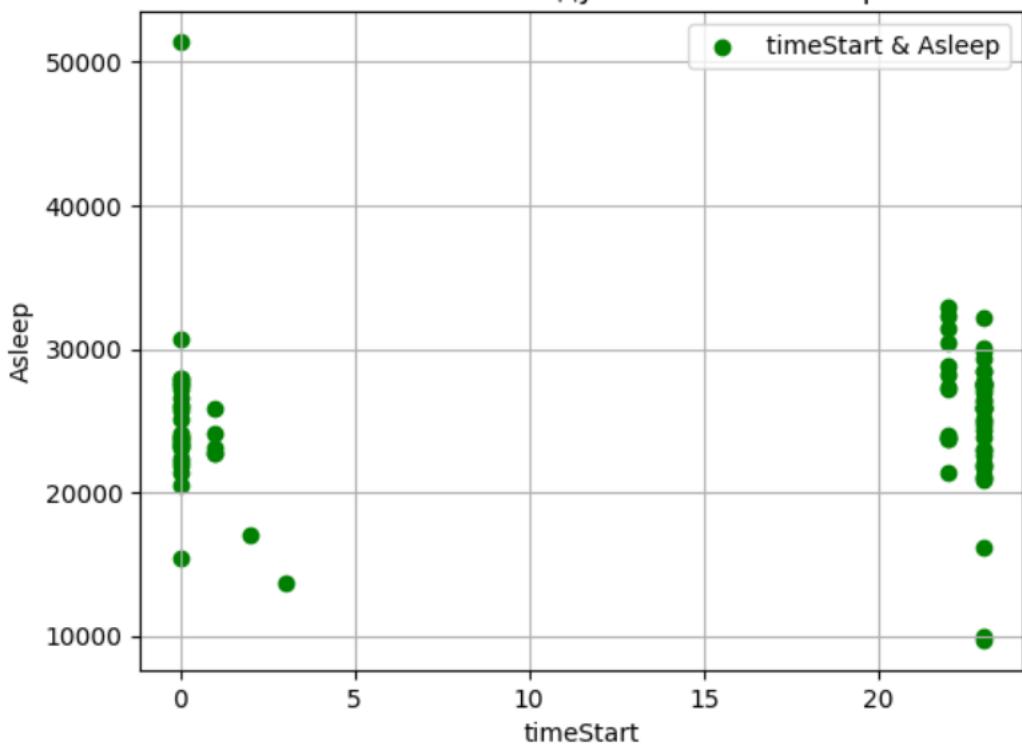
Диаграмма размаха для stagesSleep



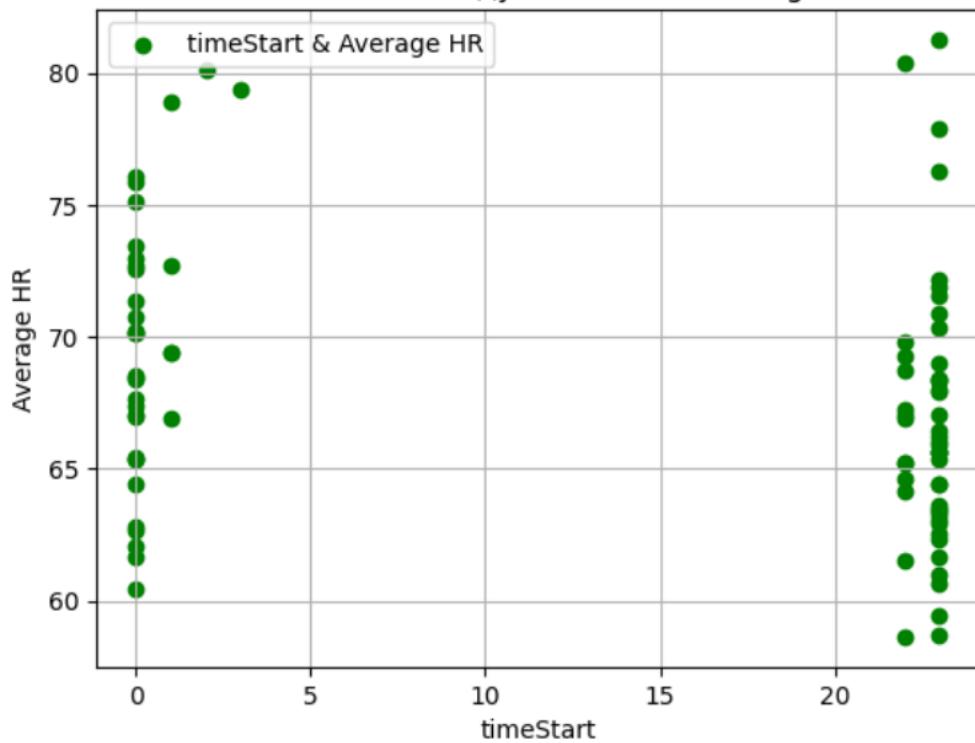
В данном случае наличие выбросов – это естественно, так как это результат работы человека над собой, а он может быть разным, поэтому их мы не трогаем.

Построим графики зависимостей одной переменной от другой.

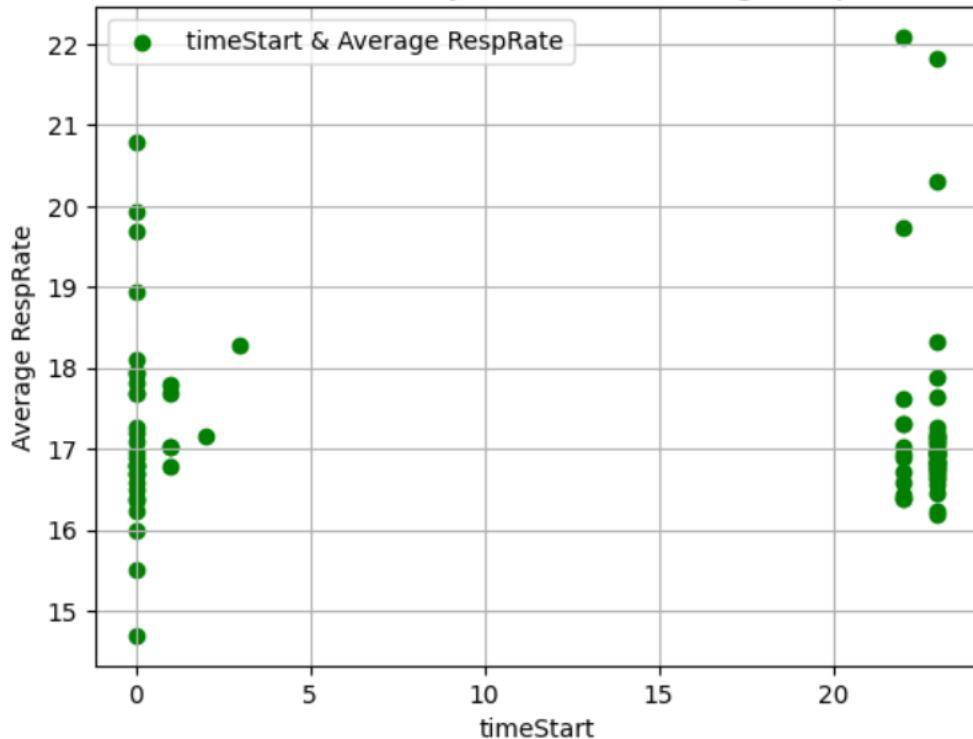
Зависимость между timeStart и Asleep



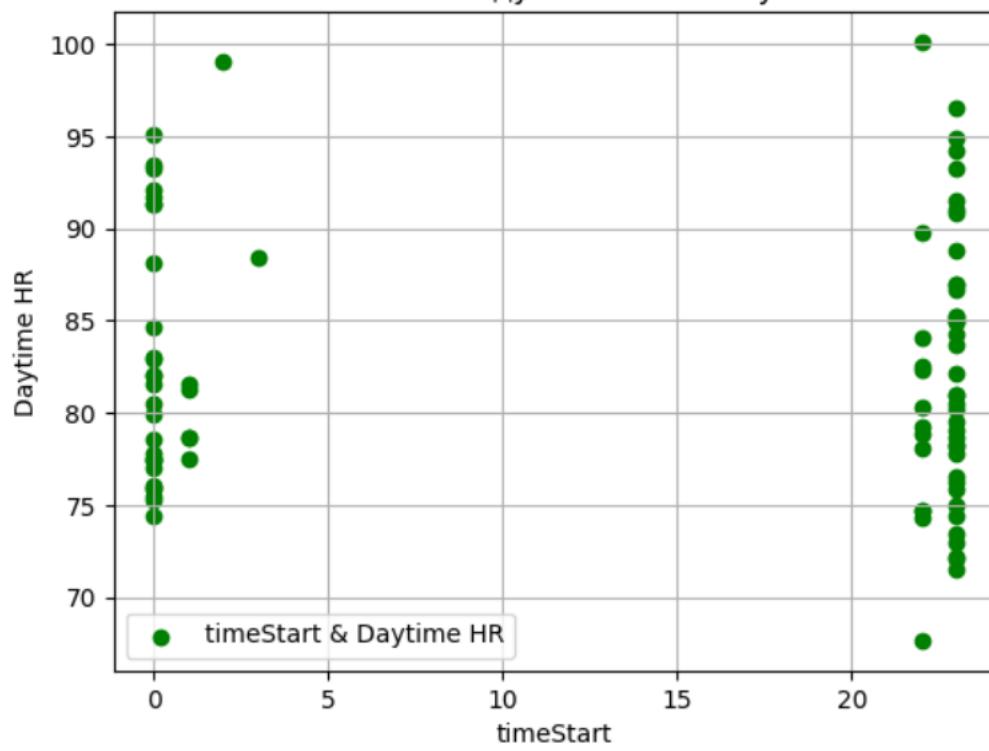
Зависимость между timeStart и Average HR



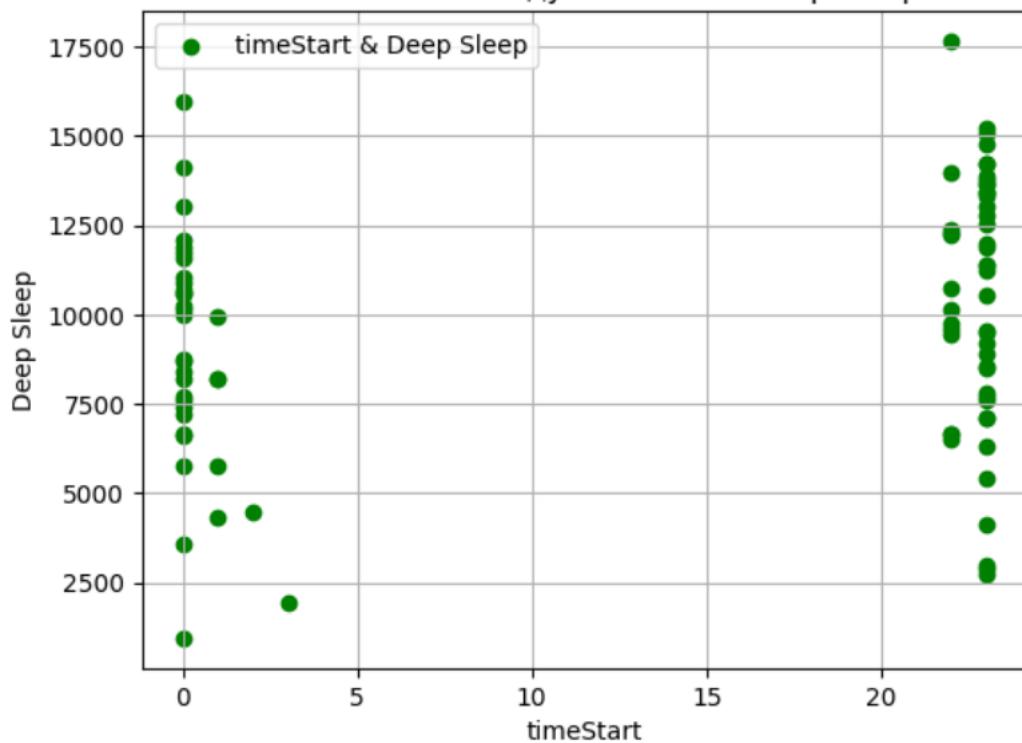
Зависимость между timeStart и Average RespRate



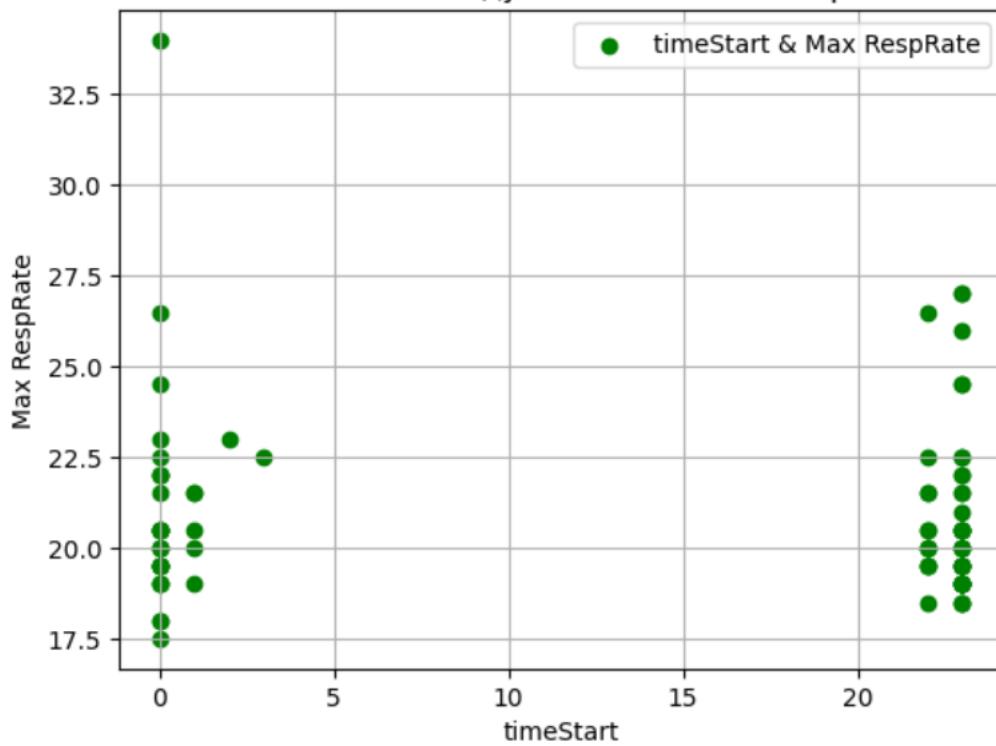
Зависимость между timeStart и Daytime HR



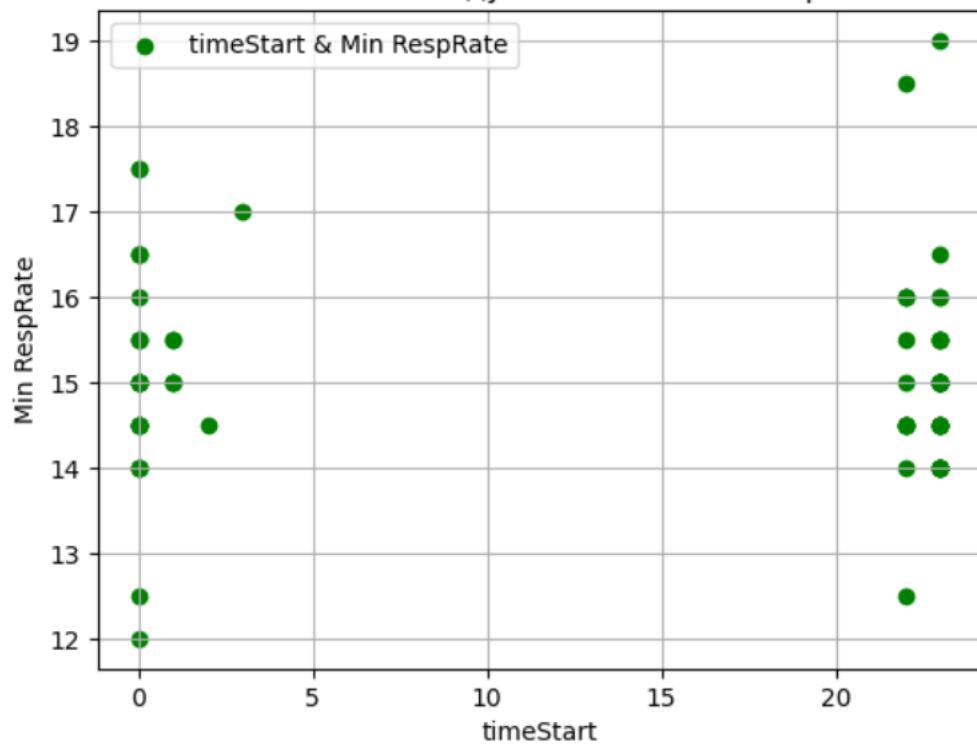
Зависимость между timeStart и Deep Sleep



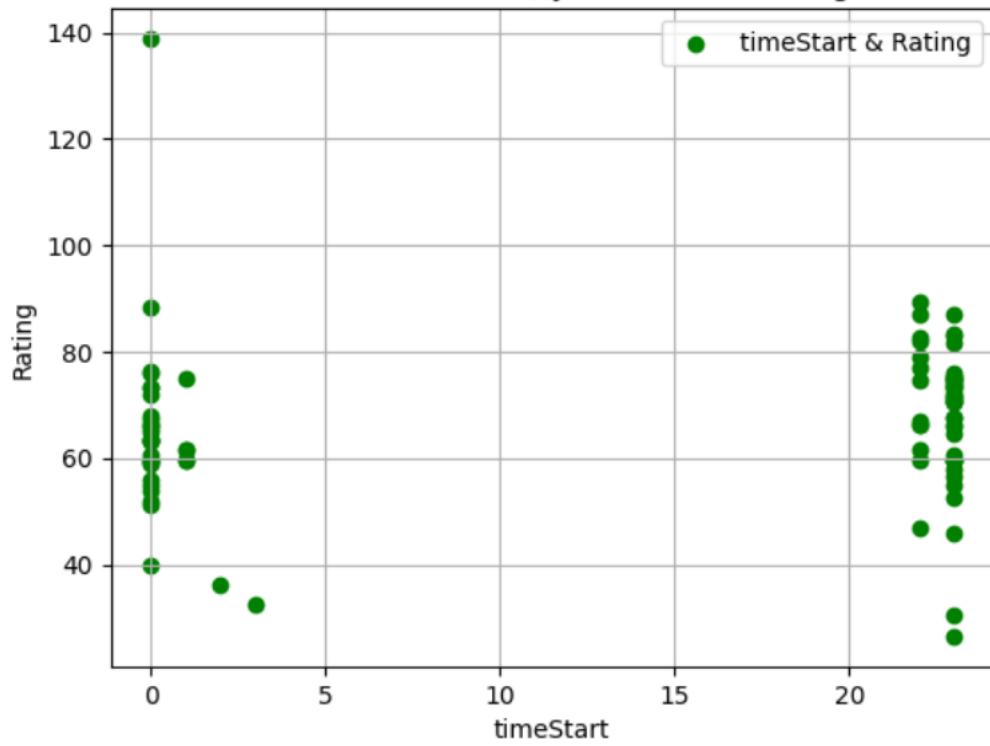
Зависимость между timeStart и Max RespRate



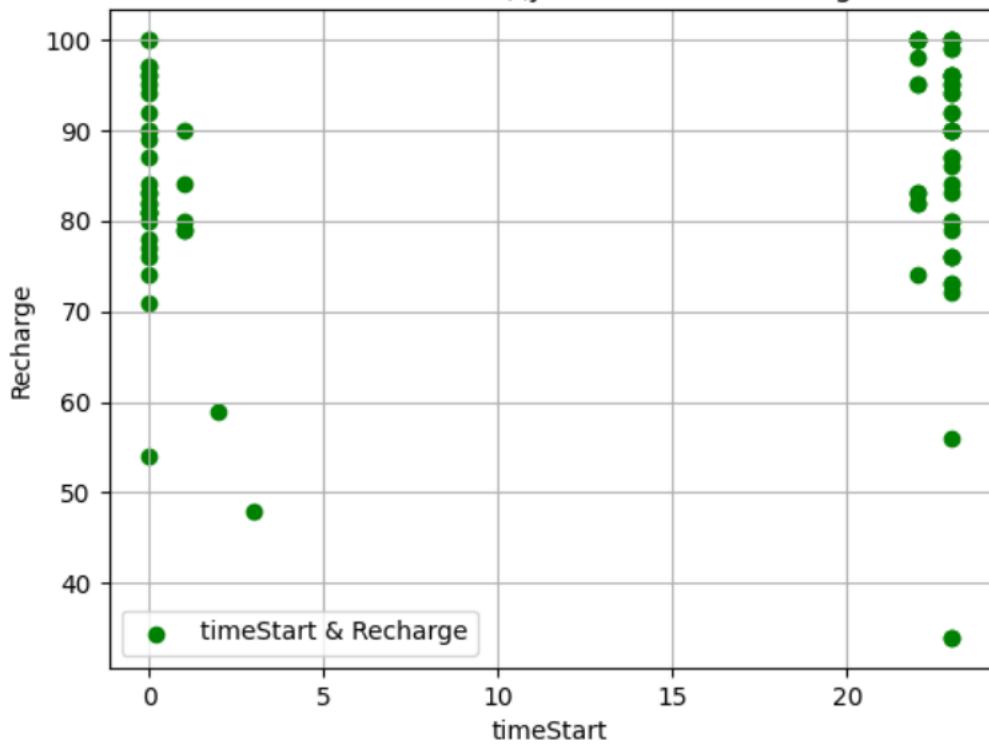
Зависимость между timeStart и Min RespRate



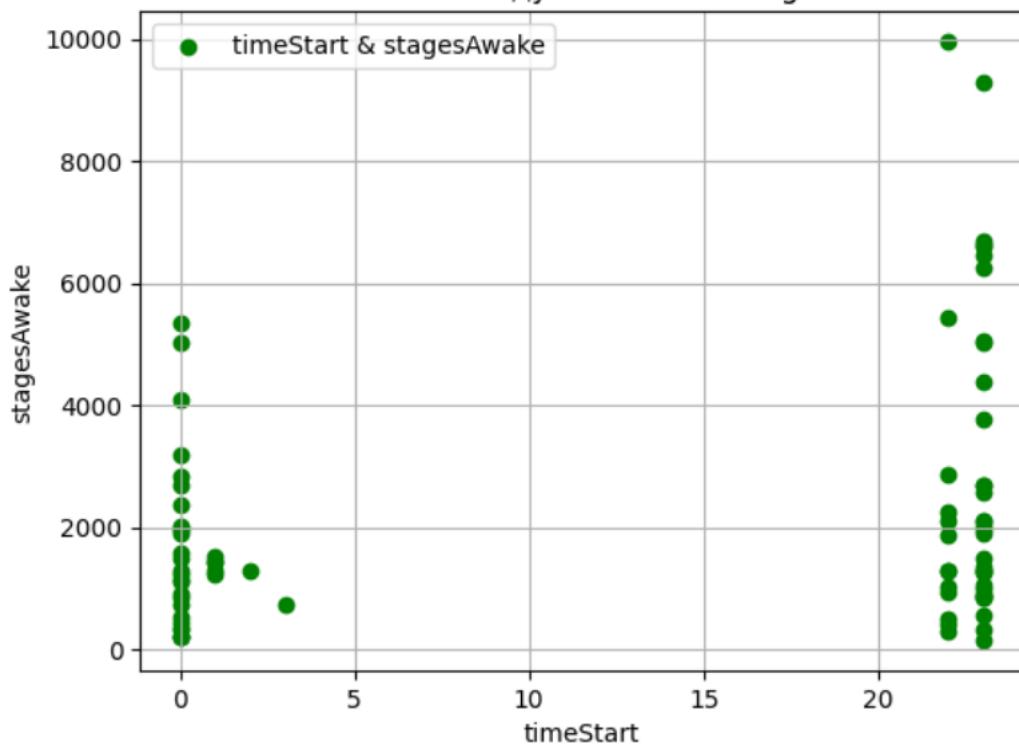
Зависимость между timeStart и Rating



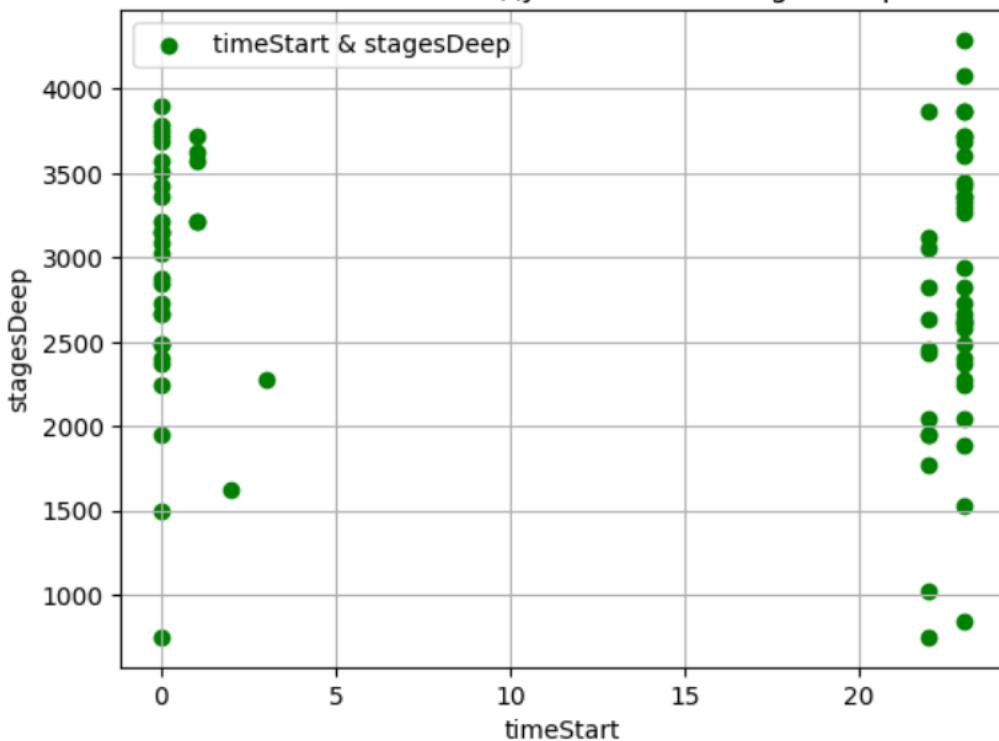
Зависимость между timeStart и Recharge



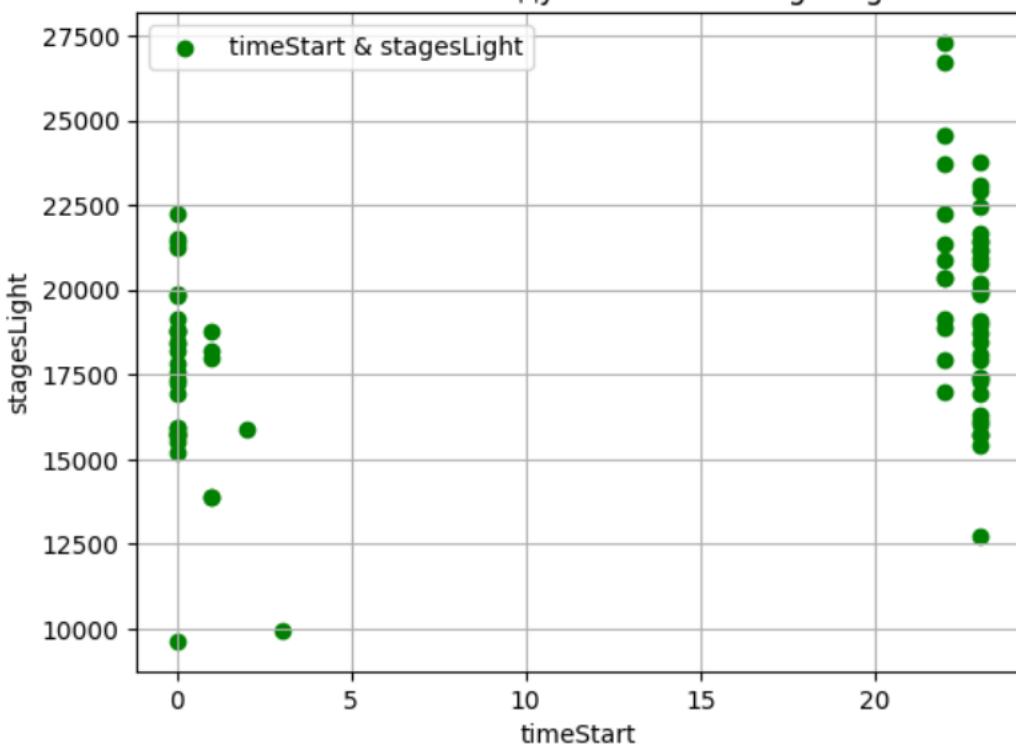
Зависимость между timeStart и stagesAwake



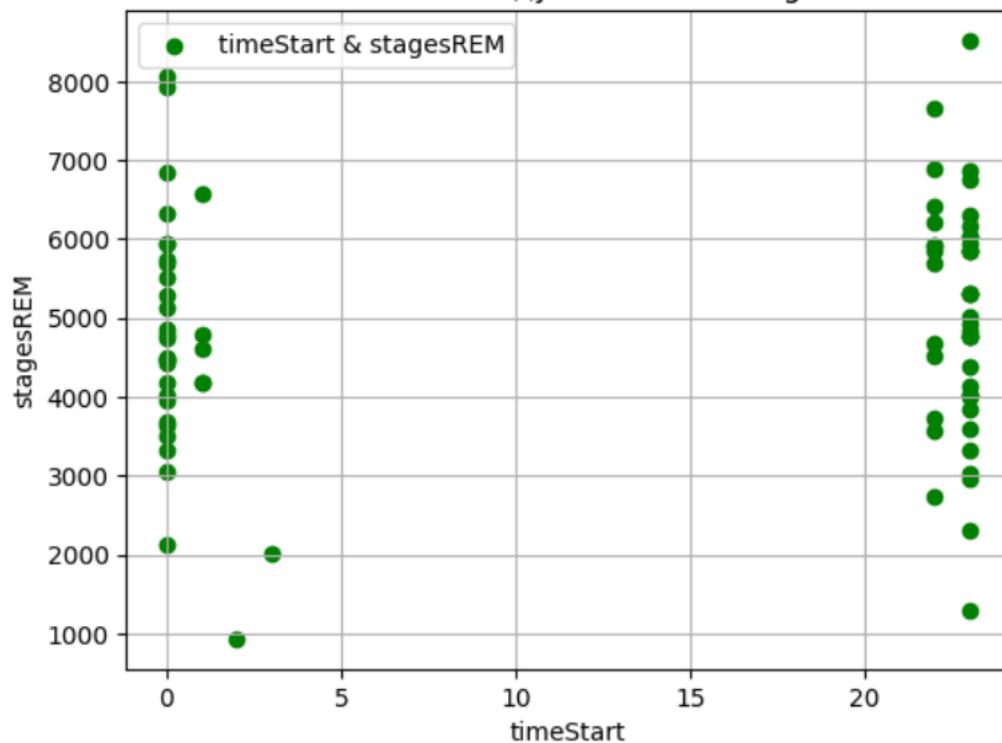
Зависимость между timeStart и stagesDeep



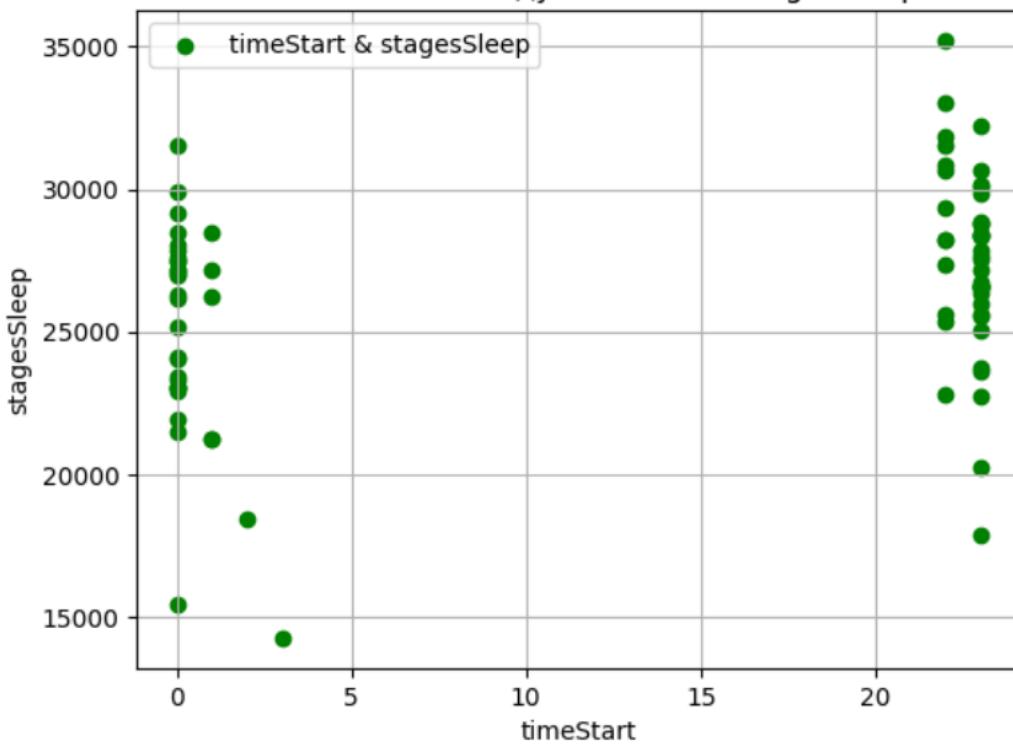
Зависимость между timeStart и stagesLight



Зависимость между timeStart и stagesREM



Зависимость между timeStart и stagesSleep



На графиках видны 2 группы. Например, между временем начала сна и четырьмя стадиями сна.

Разбор df_2

	health_kit_id	HKTTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value	...	DeepSleep	MaxRespRate	MinRespRate
0	ABCD13F9-CAB2-42E8-BEFB-63C855FC3307	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	2024-05-07T08:25:29+0200	2024-05-07T00:22:00+0200	in_bed	...	NaN	NaN	N
4	101079AB-EC9E-483C-9429-6762D23A4271	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.3.0	iPhone15,2	iPhone Аннушка	17.3	2024-02-01T05:41:17+0100	2024-01-31T22:25:00+0100	in_bed	...	NaN	NaN	N
5	9BCE117A-E0F7-4F2B-8BF9-4AADD3BA6453	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	2024-05-10T07:47:18+0200	2024-05-09T23:52:07+0200	in_bed	...	NaN	NaN	N
7	1F07E427-387C-4C2C-BF17-3EF3B1BA3BDD	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	2024-06-01T08:27:00+0200	2024-05-31T23:54:06+0200	in_bed	...	NaN	NaN	N
8	630D92DB-E3E0-4BBA-96B8-F4F4A0778365	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.1.2	iPhone15,2	iPhone Аннушка	17.1.2	2024-01-31T06:52:43+0100	2024-01-30T23:44:42+0100	in_bed	...	NaN	NaN	N

5 rows x 24 columns

Количество

строк:

148

Пропущенные значения:

	0
health_kit_id	0
HKTimeZone	0
bundleIdentifier	0
operatingSystemVersion	0
productType	0
sourceName	0
version	0
timeEnd	0
timeStart	0
value	0
Asleep	148
Average HR	148
Average RespRate	148
Daytime HR	148
Deep Sleep	148
Max RespRate	148
Min RespRate	148
Rating	148
Recharge	148
stagesAwake	148
stagesDeep	148
stagesLight	148
stagesREM	148
stagesSleep	148

Удалим признаки с пропущенными значениями, равными количеству столбцов. То есть признаки: Asleep, Average HR, Average RespRate, Daytime HR, Deep Sleep, Max RespRate, Min RespRate, Rating, Recharge, stagesAwake, stagesDeep, stagesLight, stagesREM, stagesSleep.

	0
health_kit_id	0
HKTimeZone	0
bundleIdentifier	0
operatingSystemVersion	0
productType	0
sourceName	0
version	0
timeEnd	0
timeStart	0
value	0

dtype: int64

В таком датасете нет пропущенных значений.

	0
health_kit_id	139
HKTimeZone	8
bundleIdentifier	4
operatingSystemVersion	13
productType	3
sourceName	4
version	15
timeEnd	140
timeStart	140
value	4

dtype: int64

Удалим ненужные признаки:

- health_kit_id - так как это идентификатор.

Переформатируем время, так как для анализа не требуется дата, а только время отхода ко сну и время подъема. И, чтобы ранжировать время, возьмем только часы (заодно переведем дату в числовой признак).

	HKTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value
0	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...		17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	8	0 in_bed
4	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...		17.3.0	iPhone15,2	iPhone Аннушка	17.3	5	22 in_bed
5	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...		17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	7	23 in_bed
7	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...		17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	8	23 in_bed
8	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...		17.1.2	iPhone15,2	iPhone Аннушка	17.1.2	6	23 in_bed

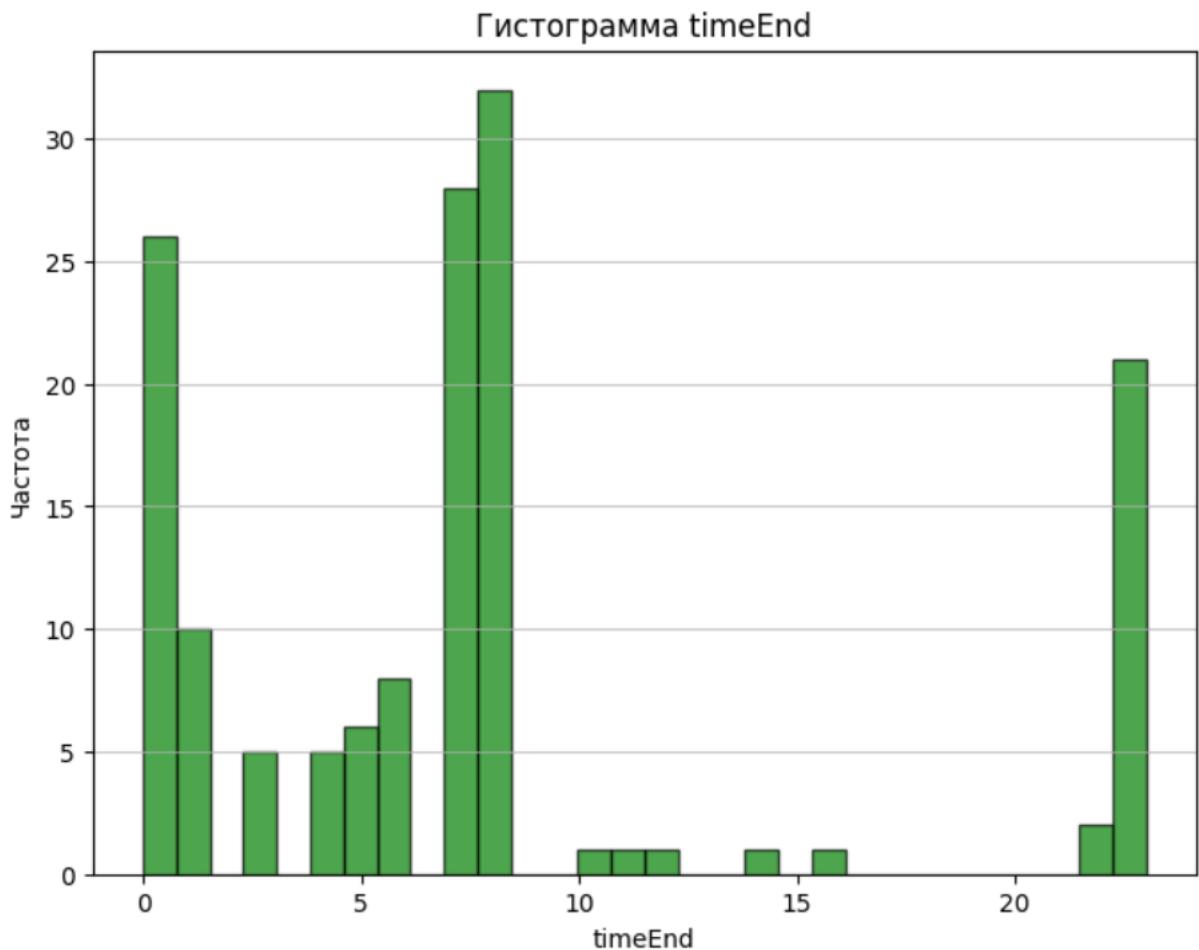
Описание числовых признаков:

	timeEnd	timeStart
count	148.000000	148.000000
mean	7.871622	16.898649
std	7.245247	9.660735
min	0.000000	0.000000
25%	3.000000	11.500000
50%	7.000000	23.000000
75%	8.000000	23.000000
max	23.000000	23.000000

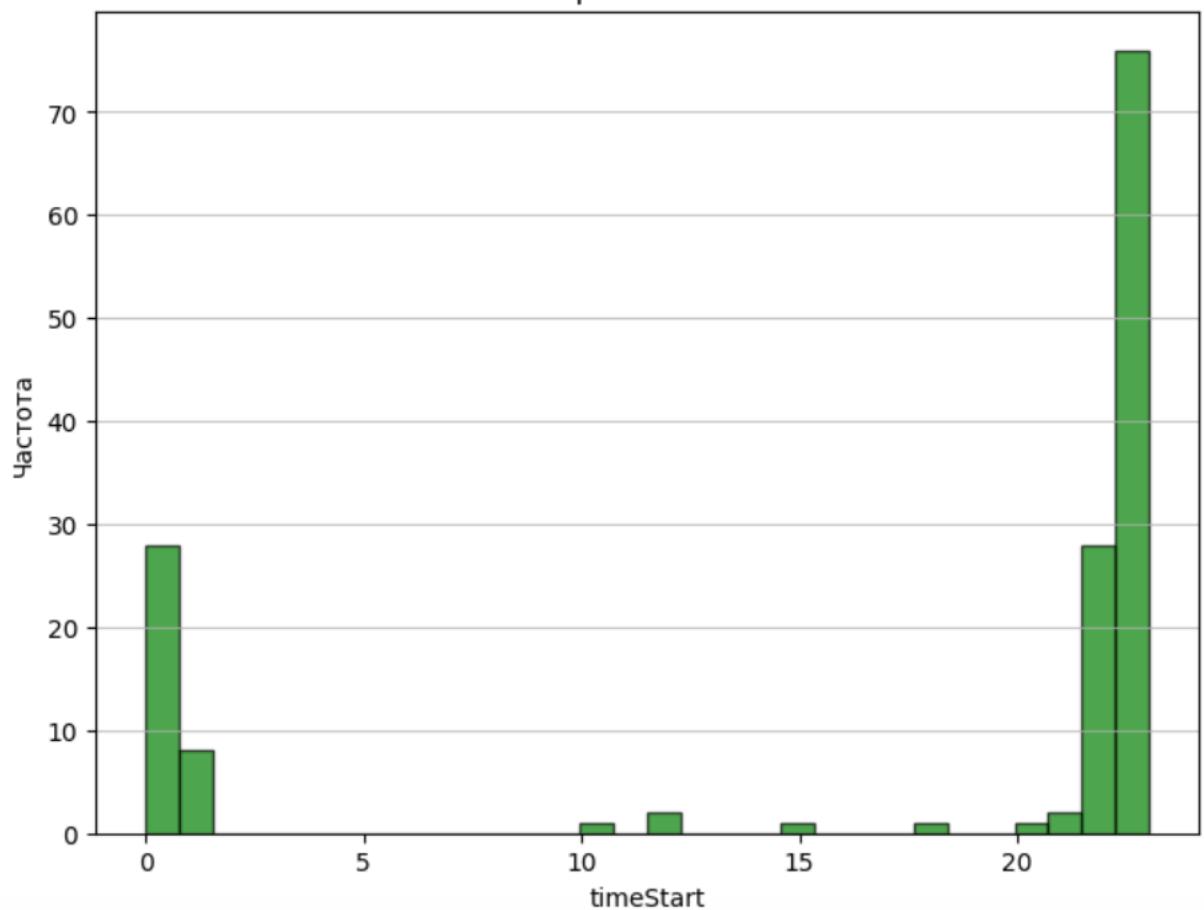
	0
HKTimeZone	object
bundleIdentifier	object
operatingSystemVersion	object
productType	object
sourceName	object
version	object
timeEnd	int64
timeStart	int64
value	object

dtype: object

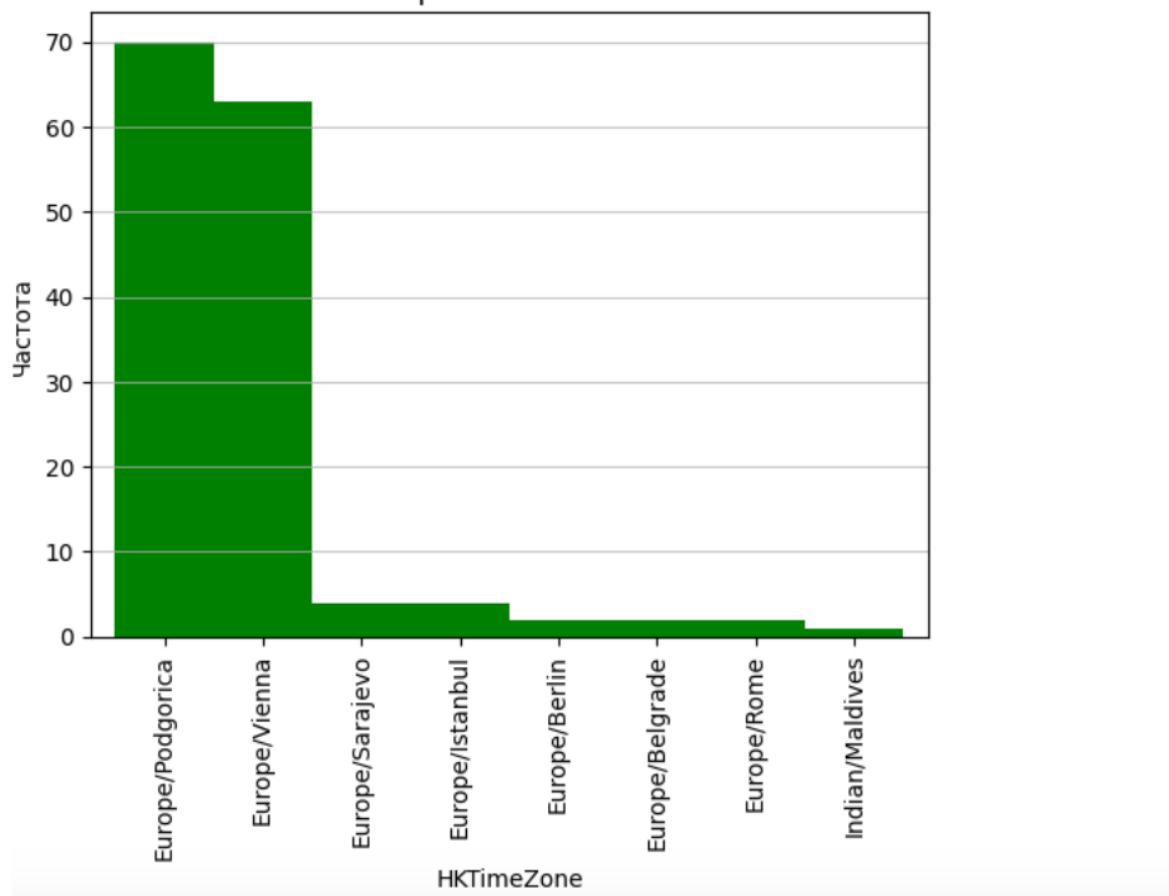
В датасете больше категориальных значений. Посмотрим на гистограммы и боксплоты.



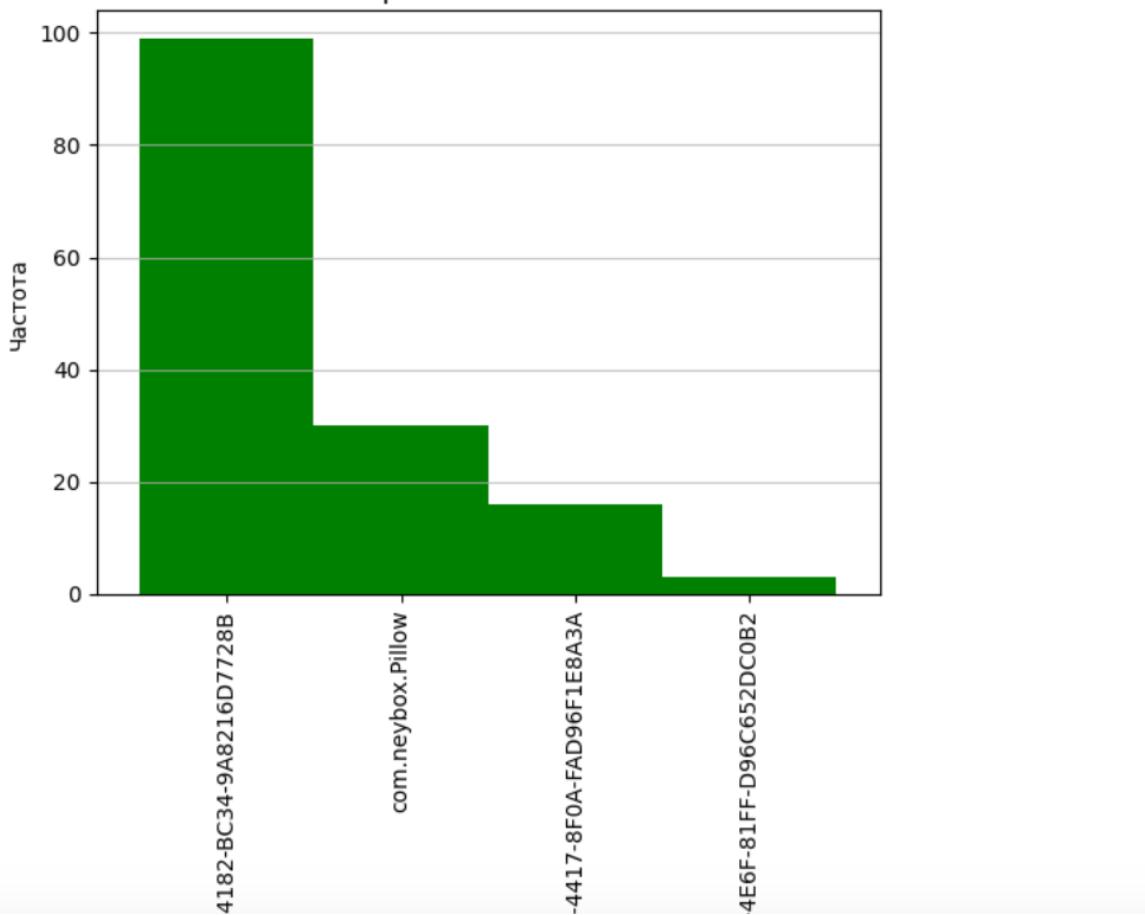
Гистограмма timeStart



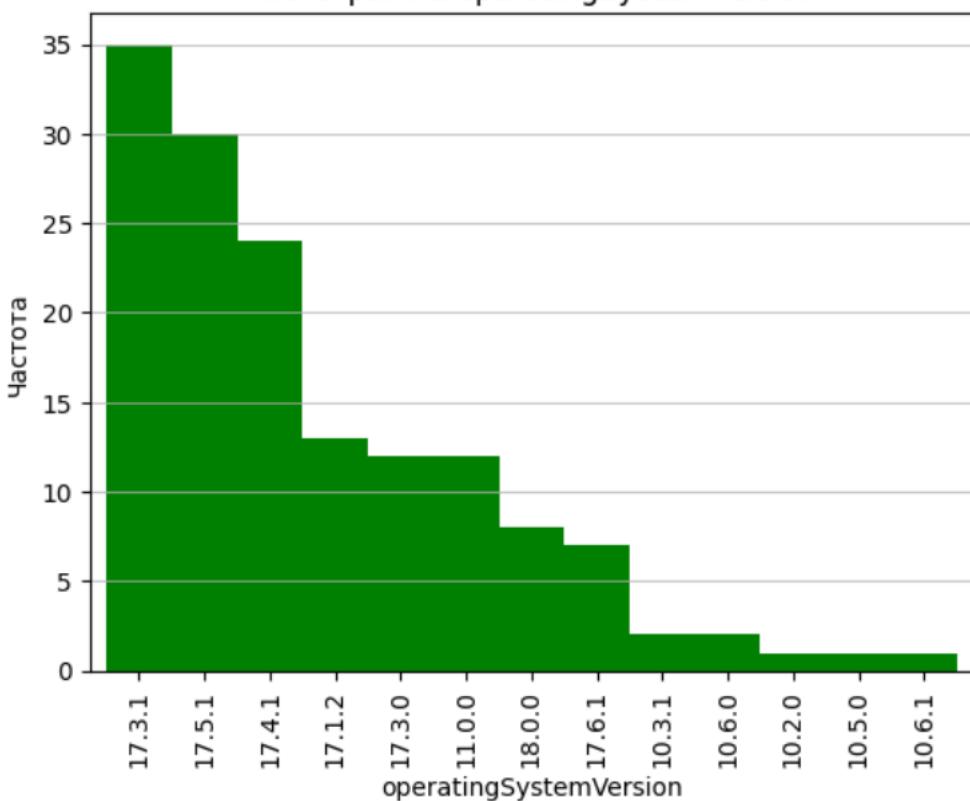
Гистограмма HTimeZone



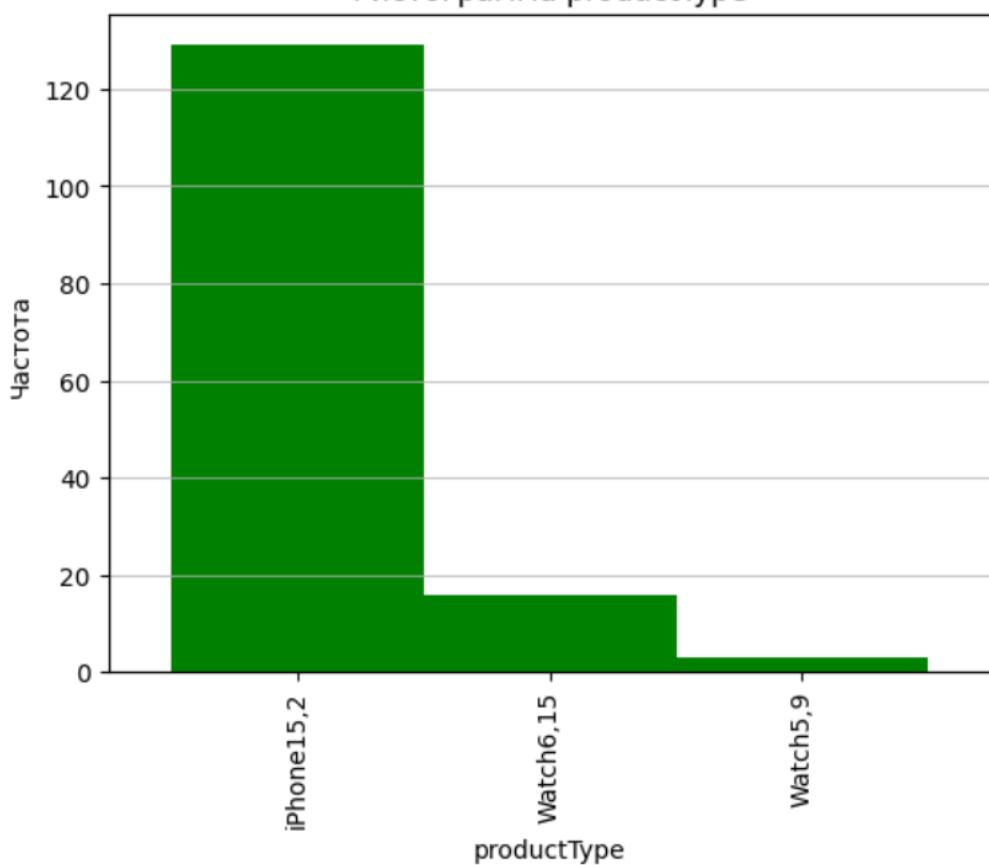
Гистограмма bundleIdentifier



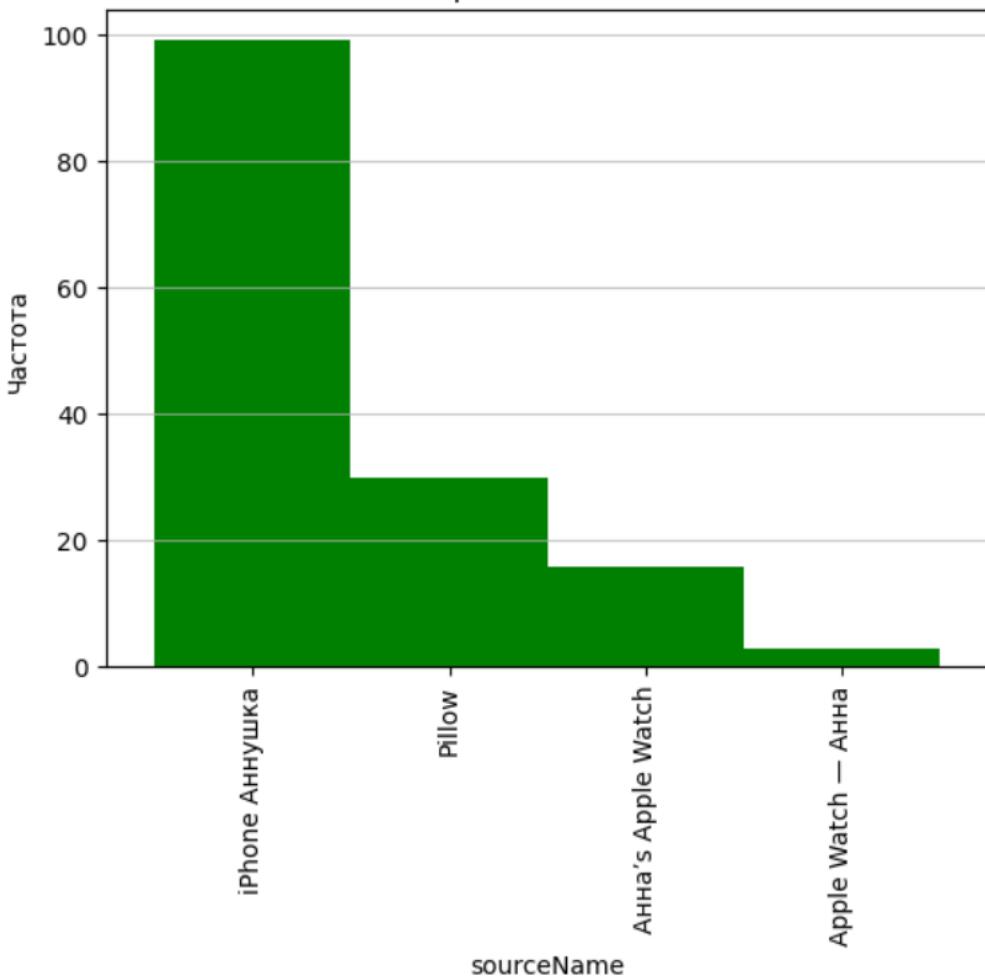
Гистограмма operatingSystemVersion



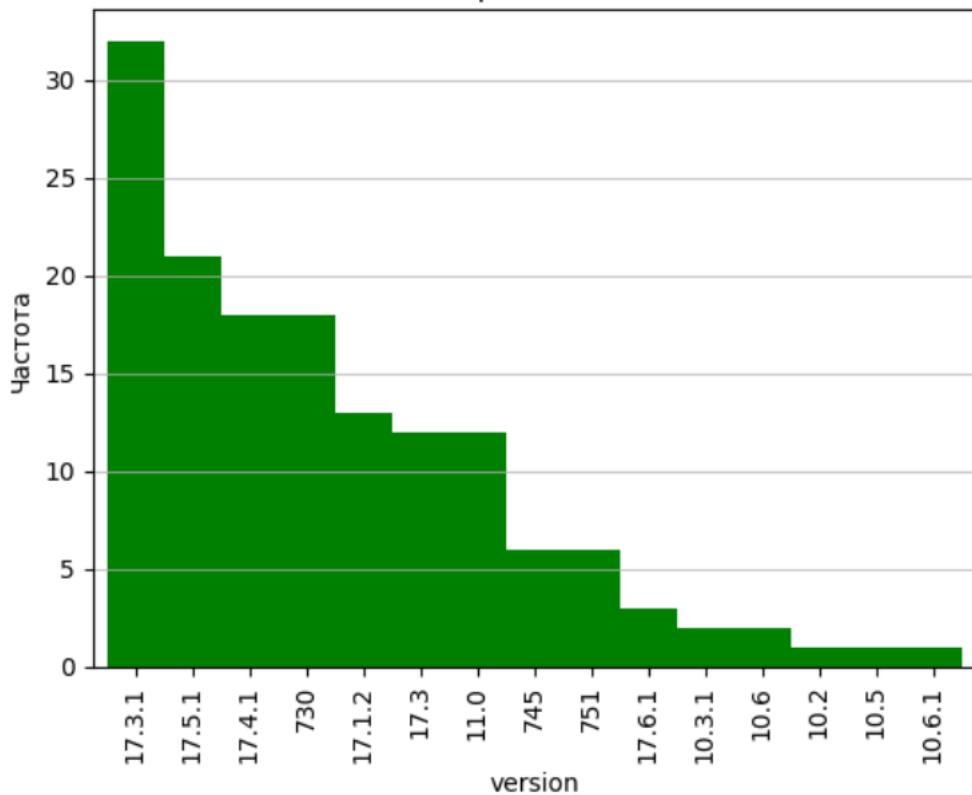
Гистограмма productType



Гистограмма sourceName



Гистограмма version



Гистограмма value

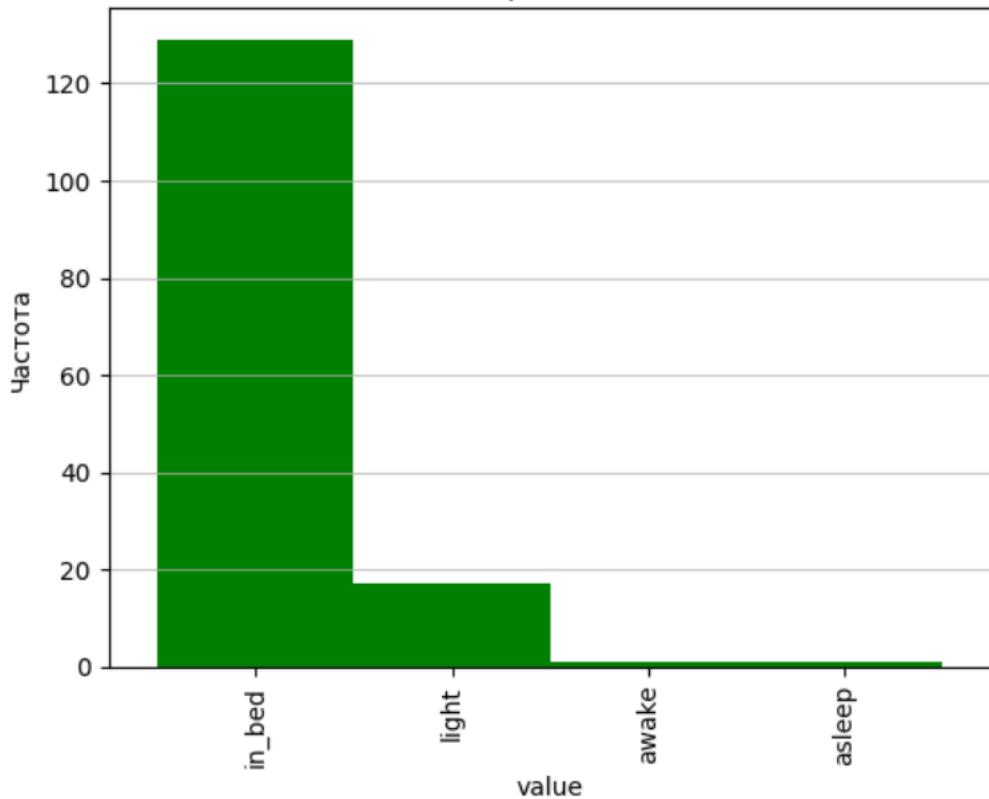


Диаграмма размаха для timeEnd

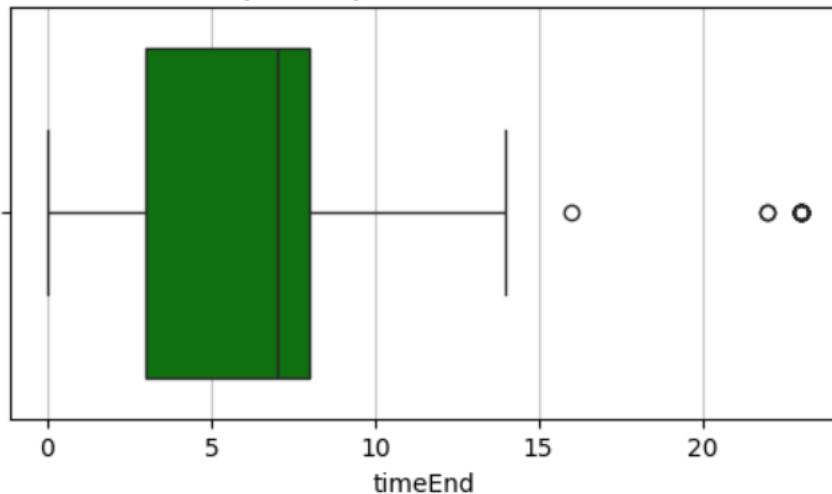
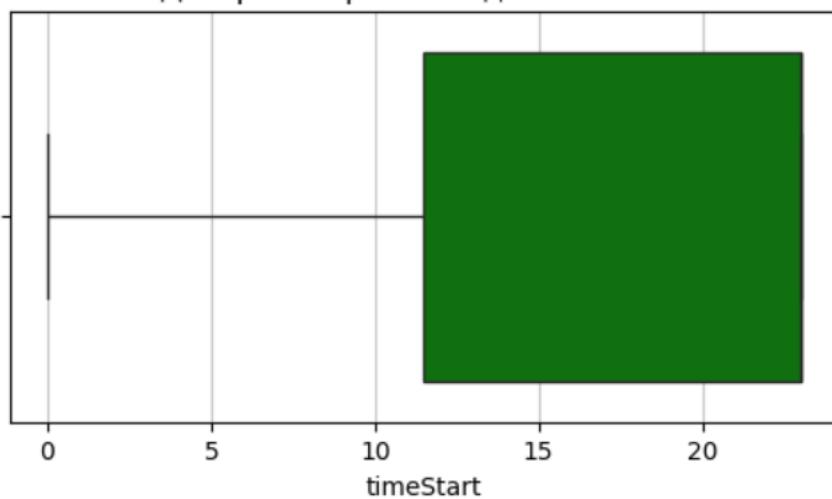
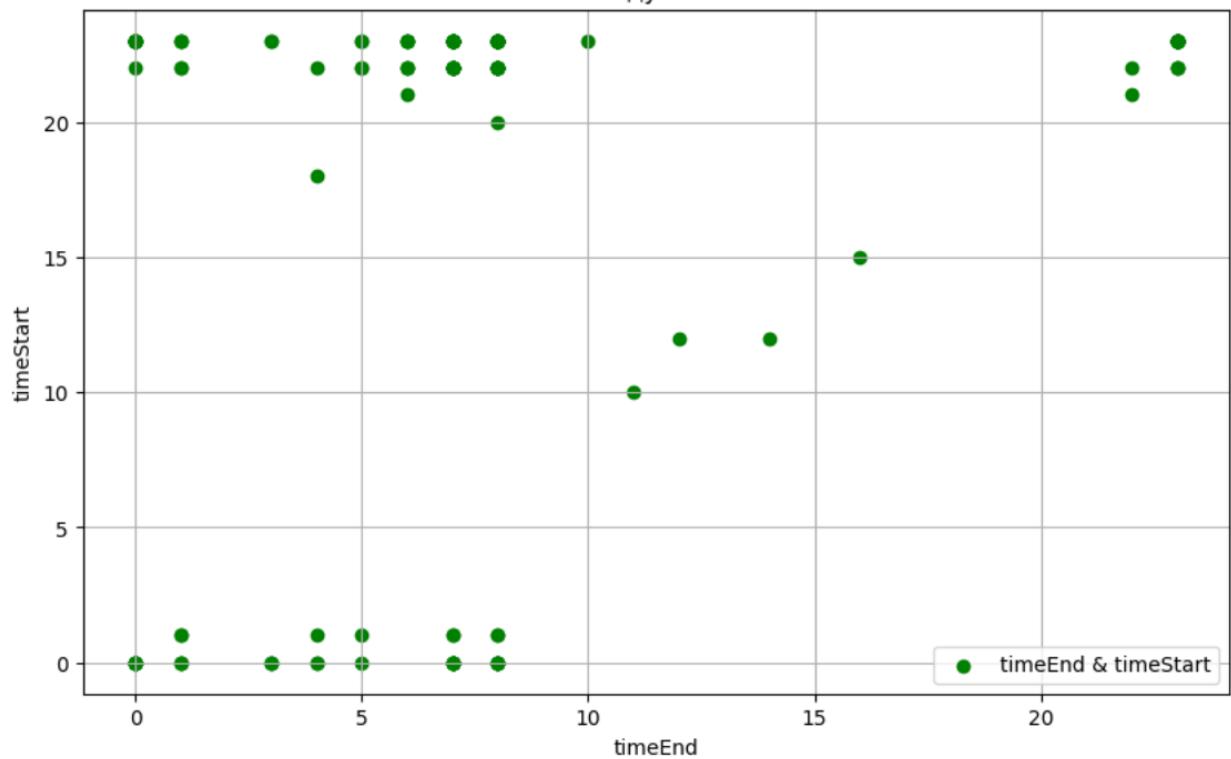


Диаграмма размаха для timeStart



Построим графики зависимостей одной переменной от другой.

Зависимость между timeEnd и timeStart



Разбор df_3

	health_kit_id	HKTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value	...	DeepSleep	MaxRespRate	MinRespRate	...
35	317C9BB0-F594-4AA4-B1CB-FB12ED5D184A	NaN	com.ouraring.oura	17.6.1	iPhone15,2	Oura	2408150708	2024-09-08T07:51:00+0200	2024-09-07T22:34:00+0200	in_bed	...	NaN	NaN	NaN	...
36	27280CDC8-5636-490E-B3A9-745D920BB2EB	NaN	com.apple.Health	18.0.0	iPhone15,2	Health	18.0	2024-11-02T07:49:00+0100	2024-11-01T22:30:00+0100	in_bed	...	NaN	NaN	NaN	...
45	99167BA8-C964-4B32-A65D-C62F261BE449	NaN	com.ouraring.oura	18.0.0	iPhone15,2	Oura	2409241358	2024-11-03T09:00:33+0100	2024-11-02T23:27:03+0100	in_bed	...	NaN	NaN	NaN	...
98	A8353BE9-61DB-4E31-8F6D-ABDB35953175	NaN	com.ouraring.oura	18.0.0	iPhone15,2	Oura	2409241358	2024-10-16T07:08:00+0200	2024-10-15T22:59:00+0200	in_bed	...	NaN	NaN	NaN	...
178	64F0B351-F1AF-4570-966D-42F4F1ED1504	NaN	com.ouraring.oura	18.0.0	iPhone15,2	Oura	2409241358	2024-10-17T07:57:02+0200	2024-10-17T00:54:02+0200	in_bed	...	NaN	NaN	NaN	...

5 rows x 24 columns

Количество строк: 6

Пропущенные значения:

	0
health_kit_id	0
HKTimeZone	6
bundleIdentifier	0
operatingSystemVersion	0
productType	0
sourceName	0
version	0
timeEnd	0
timeStart	0
value	0
Asleep	6
Average HR	6
Average RespRate	6
Daytime HR	6
Deep Sleep	6
Max RespRate	6
Min RespRate	6
Rating	6
Recharge	6
stagesAwake	6
stagesDeep	6
stagesLight	6
stagesREM	6
stagesSleep	6

Удалим столбцы, в которых количество выбросов совпадает с количеством строк.

	0
health_kit_id	6
bundleIdentifier	2
operatingSystemVersion	2
productType	1
sourceName	2
version	3
timeEnd	6
timeStart	6
value	1

dtype: int64

Удалим идентификатор и закодируем дату и время.

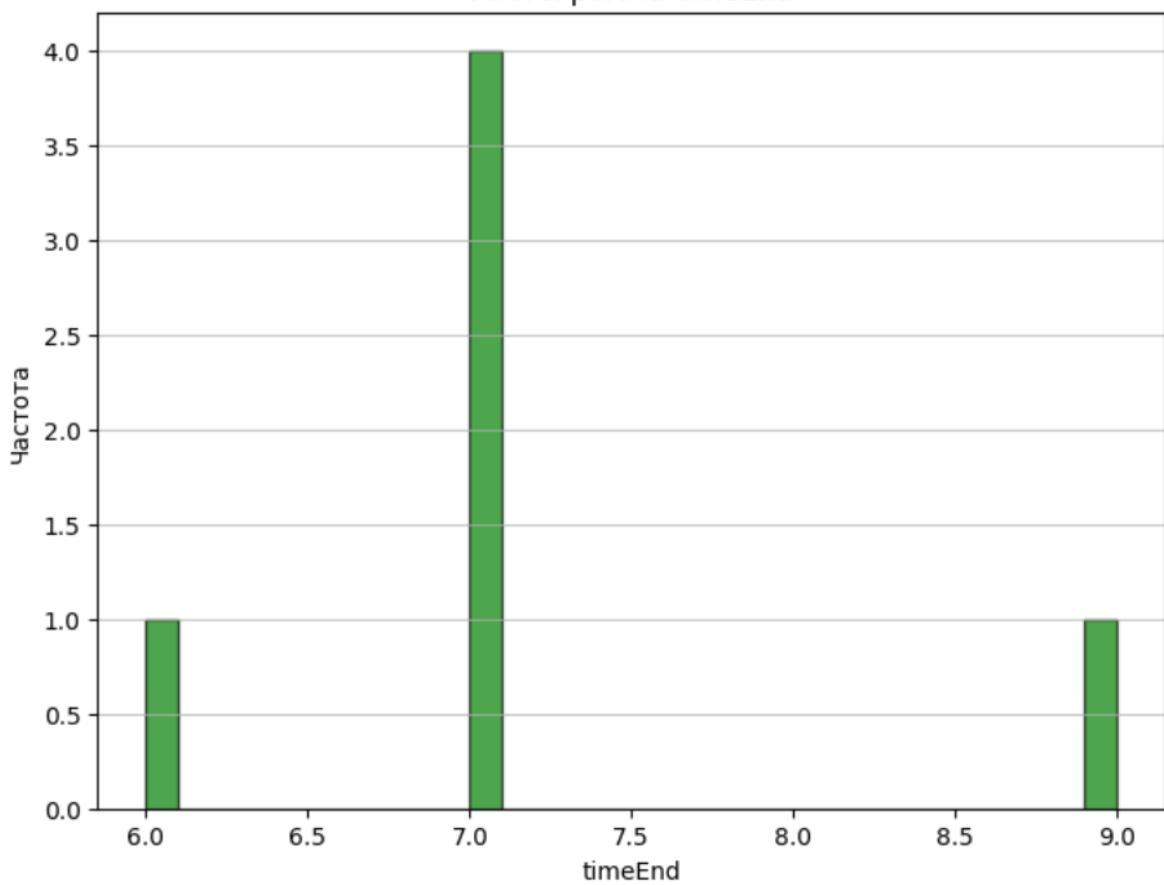
	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value
35	com.ouraring.oura	17.6.1	iPhone15,2	Oura	2408150708	7	22	in_bed
36	com.apple.Health	18.0.0	iPhone15,2	Health	18.0	7	22	in_bed
45	com.ouraring.oura	18.0.0	iPhone15,2	Oura	2409241358	9	23	in_bed
98	com.ouraring.oura	18.0.0	iPhone15,2	Oura	2409241358	7	22	in_bed
178	com.ouraring.oura	18.0.0	iPhone15,2	Oura	2409241358	7	0	in_bed

Удалим и уникальные значения.

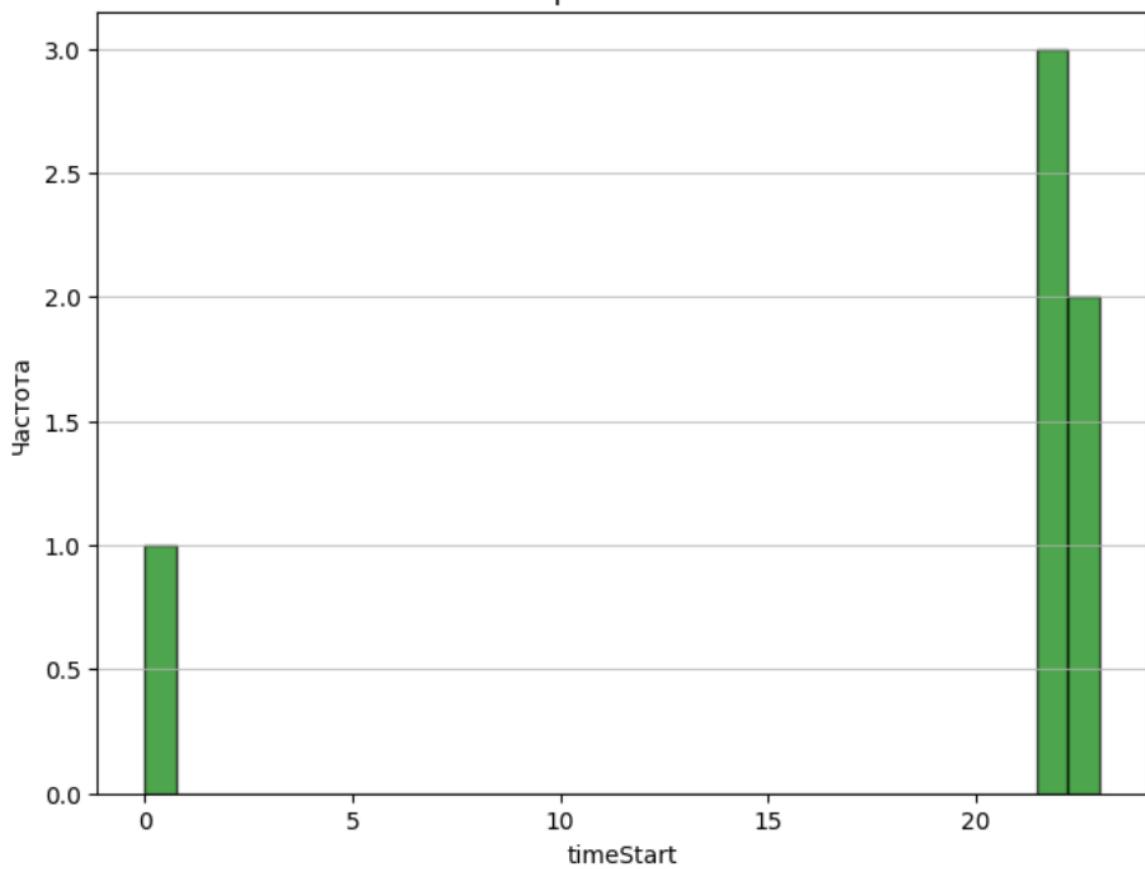
	bundleIdentifier	operatingSystemVersion	sourceName	version	timeEnd	timeStart
35	com.ouraring.oura	17.6.1	Oura	2408150708	7	22
36	com.apple.Health	18.0.0	Health	18.0	7	22
45	com.ouraring.oura	18.0.0	Oura	2409241358	9	23
98	com.ouraring.oura	18.0.0	Oura	2409241358	7	22
178	com.ouraring.oura	18.0.0	Oura	2409241358	7	0

В данном датафрейме описание и боксплоты совпадают со вторым датафреймом, поэтому не будем их строить, посмотрим только на гистограммы.

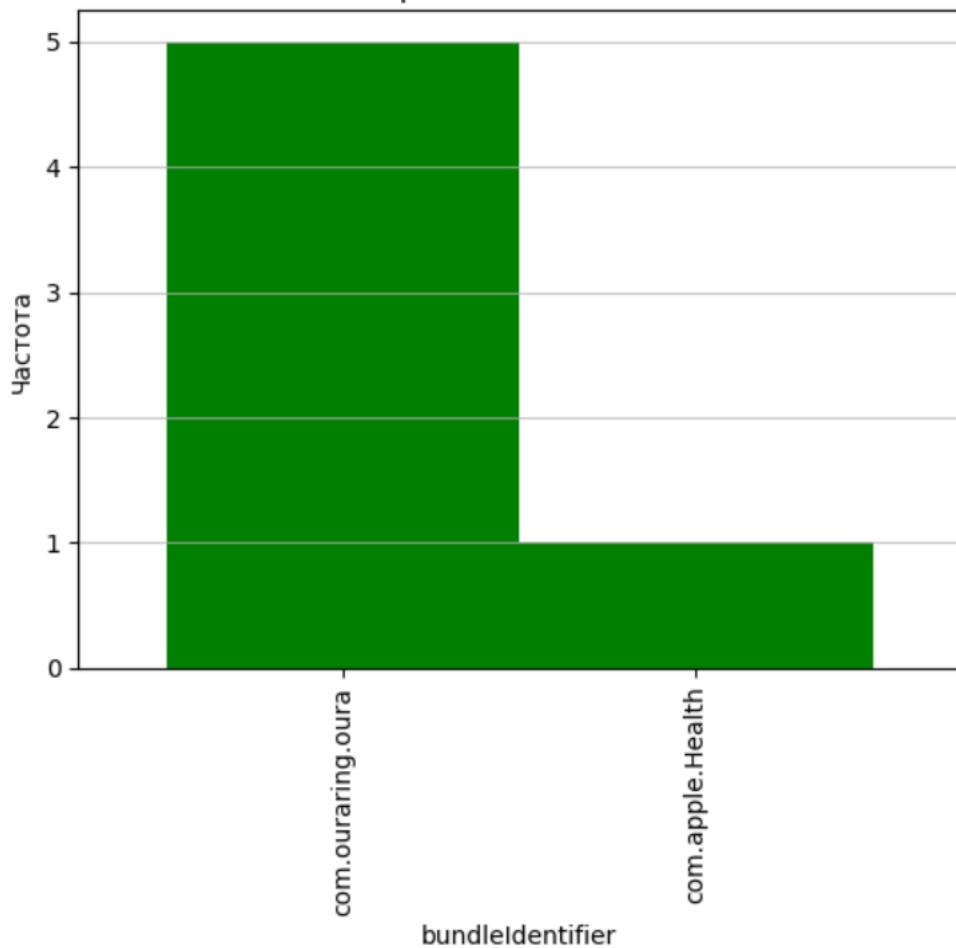
Гистограмма timeEnd



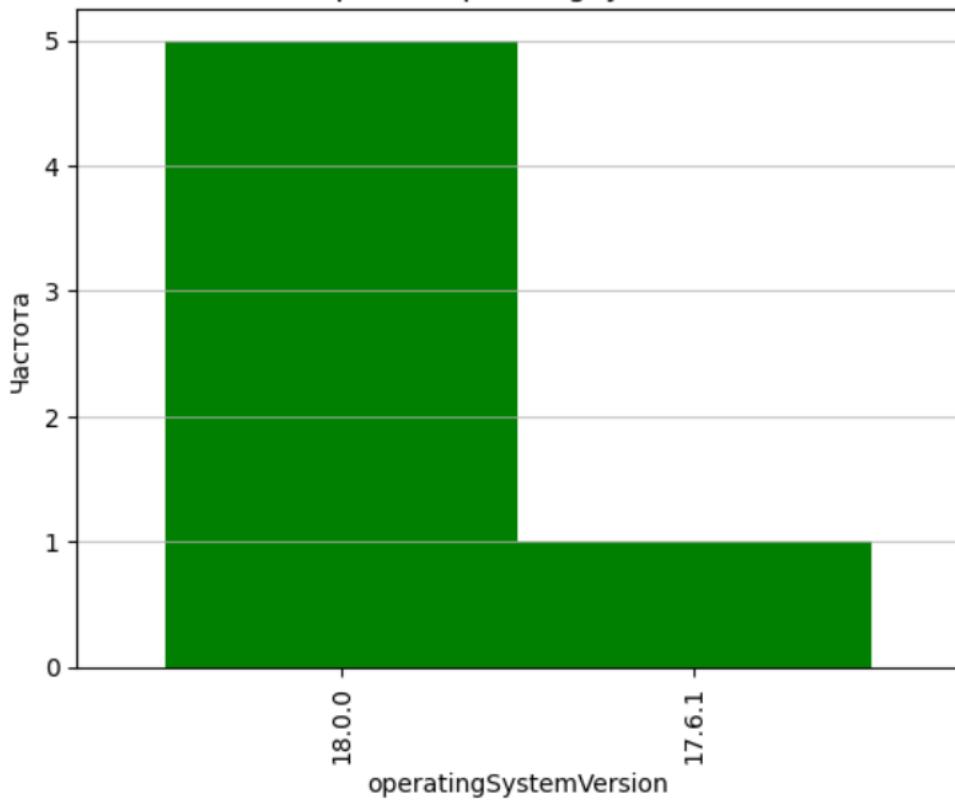
Гистограмма timeStart



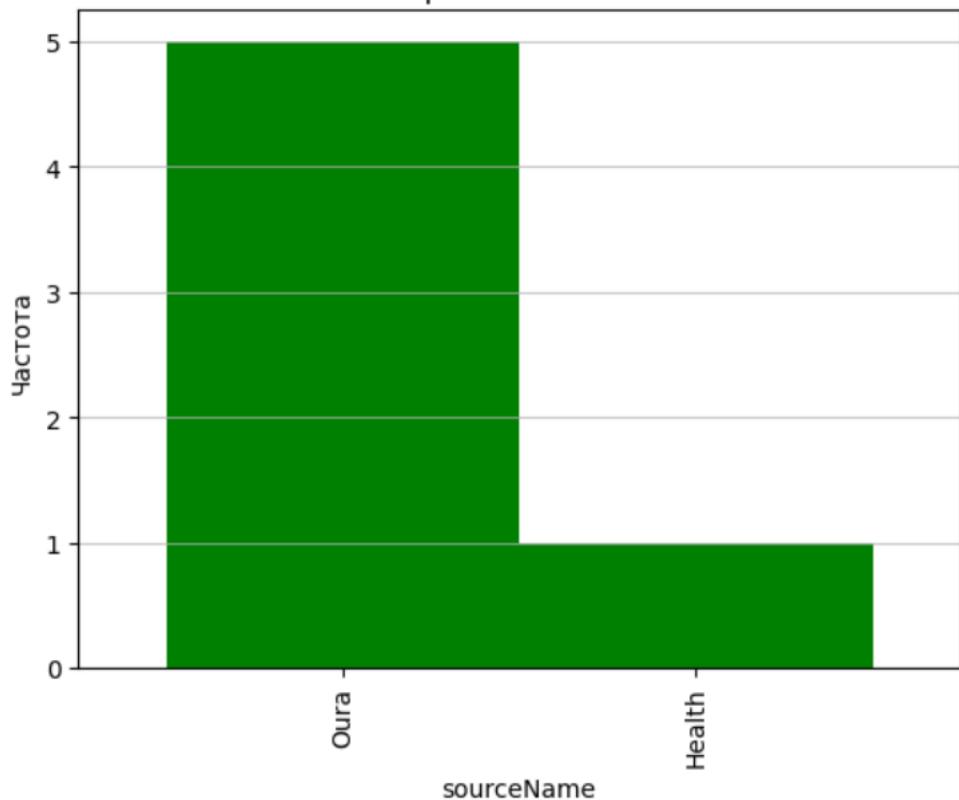
Гистограмма bundleIdentifier



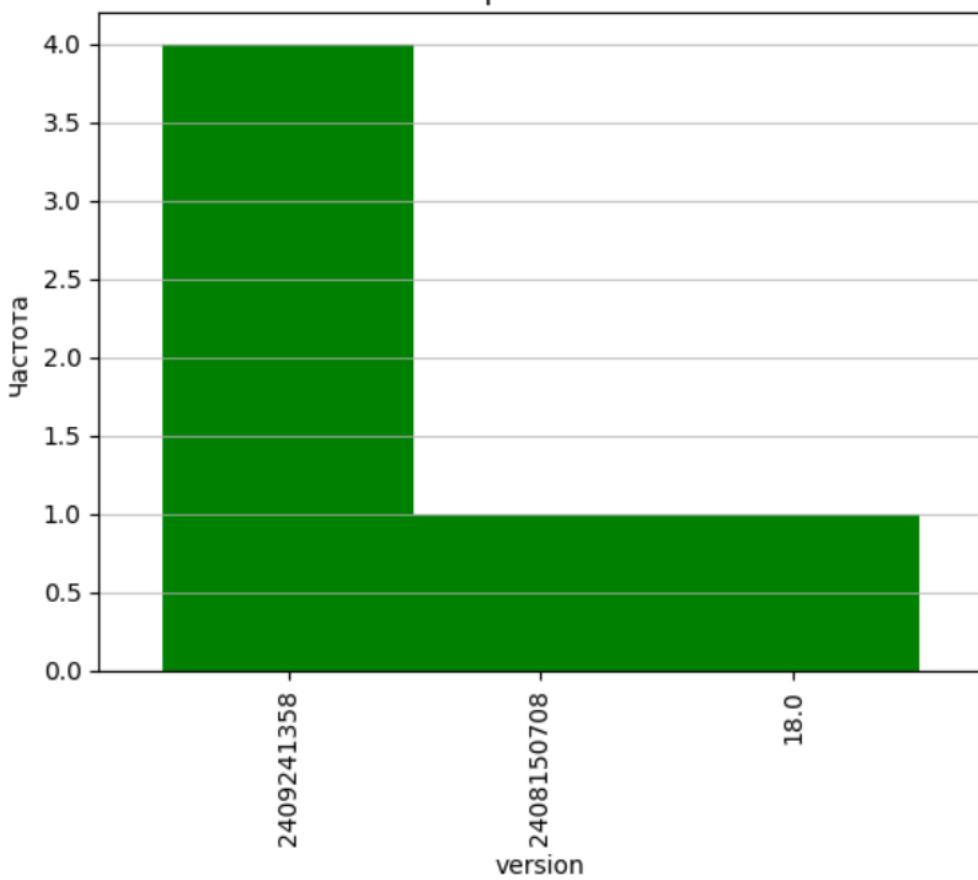
Гистограмма operatingSystemVersion



Гистограмма sourceName



Гистограмма version



Выводы к разделу 1

Разделение на три датафрейма привело к следующим результатам. Предположение следующее: было несколько исследований, которые объединили в единый датасет. Мы их разделили. Анализ трех датафреймов показал, что результаты первого исследования позволяют классифицировать и проанализировать данные. Результаты второго и третьего же - нет. Они больше показывает, какие устройства записали данные, а второй дополнительно и часовой пояс, в котором тестировался исследуемый.

Часть 2. Анализ источников

Посмотрим на итоговые датафреймы еще раз:

	operatingSystemVersion	version	timeEnd	timeStart	Asleep	Average HR	Average RespRate	Daytime HR	Deep Sleep	Max RespRate	Min RespRate	Rating	Recharge	stagesAwake	stagesDeep	stagesLight	stagesSleep
1	18.0.0	6.11.21	7	23	25020.0	63.50	16.7564	84.27	12510.0	22.5	15.0	70.73	87.0	870.0	2670.0	18750.0	361
2	18.0.0	6.10.30	6	0	21360.0	73.44	19.9405	93.23	8738.0	23.0	17.5	54.86	74.0	5340.0	1500.0	15510.0	451
3	17.5.1	6.10.30	8	23	28500.0	66.26	16.8250	79.53	9500.0	18.5	15.0	75.88	99.0	2010.0	2250.0	23760.0	41
6	17.5.1	6.10.30	7	23	27660.0	68.31	16.7167	81.00	9220.0	19.5	14.0	73.31	96.0	1290.0	2640.0	21390.0	48
15	17.6.1	6.10.30	6	22	23880.0	68.74	16.3942	84.09	9729.0	20.0	14.5	61.70	83.0	2100.0	3870.0	17910.0	35

Здесь нет источников, так как источник единственный на выборку - AutoSleep. При разборе данного датафрейма мы удалили столбец с этой информацией за ненадобностью.

AutoSleep считывает параметры: Asleep, Average HR, Average RespRate, Daytime HR, Deep Sleep, Max RespRate, Min RespRate, Rating, Recharge, stagesAwake, stagesDeep, stagesLight, stagesREM, stagesSleep.

1. **Asleep** – Общее время сна. Это показатель, который отражает, сколько времени вы провели в состоянии сна.
2. **Average HR (Average Heart Rate)** – Средний пульс. Это среднее значение частоты сердечных сокращений за время сна.
3. **Average RespRate (Average Respiratory Rate)** – Средняя частота дыхания. Указывает, сколько дыхательных циклов происходило за минуту в среднем во время сна.
4. **Daytime HR** – Пульс в течение дня. Отражает среднюю частоту сердечных сокращений в период бодрствования.
5. **Deep Sleep** – Глубокий сон. Это показатель времени, проведенного в фазе глубокого сна, которая важна для физического восстановления.
6. **Max RespRate (Maximum Respiratory Rate)** – Максимальная частота дыхания. Указывает на максимальное значение частоты дыхания, зафиксированное во время сна.
7. **Min RespRate (Minimum Respiratory Rate)** – Минимальная частота дыхания. Показывает минимальное значение частоты дыхания за время сна.

8. **Rating** – Оценка сна. Общая оценка качества сна на основе различных параметров.
9. **Recharge** – Восстановление. Это значение может указывать на степень восстановления организма во время сна.
10. **stagesAwake** – Время бодрствования. Отражает количество времени, проведенного в состоянии бодрствования во время предполагаемого времени сна.
11. **stagesDeep** – Время глубокого сна. Указывает на количество времени, проведенного в глубоком сне.
12. **stagesLight** – Время легкого сна. Отражает количество времени, проведенного в легком (менее глубоком) сне.
13. **stagesREM** – Время быстрого сна (REM-сна). Указывает на количество времени, проведенного в фазе быстрого сна, связанной с сновидениями и когнитивными процессами.
14. **stagesSleep** – Общее время сна. В этом контексте это значение может относиться к общему количеству времени, проведенного в разных стадиях сна.

Эти параметры помогают оценить качество и структуру сна, а также выявить возможные проблемы, связанные с его нарушениями.

Имея представление о втором получившемся датафрейме (с другими источниками), можно сказать, что AutoSleep делает более подробные замеры для анализа сна. Однако, для чистоты эксперимента разберем источники остальных датафреймов.

	HKTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value
0	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	8	0	in_bed
4	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.3.0	iPhone15,2	iPhone Аннушка	17.3	5	22	in_bed
5	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	7	23	in_bed
7	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	8	23	in_bed
8	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.1.2	iPhone15,2	iPhone Аннушка	17.1.2	6	23	in_bed

Здесь 4 источника: 'Apple Watch — Анна', 'Pillow', 'iPhone Аннушка', 'Анна's Apple Watch'.

	HKTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value
41	Europe/Podgorica	com.apple.health.A4F74977-C6D7-4E6F-81FF-D96C6...	10.3.1	Watch5,9	Apple Watch — Анна	10.3.1	1	0	light
64	Europe/Podgorica	com.apple.health.A4F74977-C6D7-4E6F-81FF-D96C6...	10.3.1	Watch5,9	Apple Watch — Анна	10.3.1	0	23	light
66	Europe/Podgorica	com.apple.health.A4F74977-C6D7-4E6F-81FF-D96C6...	10.2.0	Watch5,9	Apple Watch — Анна	10.2	1	1	light

	HKTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value
9	Europe/Podgorica	com.neybox.Pillow	17.3.1	iPhone15,2	Pillow	730	3	0	in_bed
12	Europe/Vienna	com.neybox.Pillow	17.4.1	iPhone15,2	Pillow	730	8	0	in_bed
17	Europe/Podgorica	com.neybox.Pillow	17.3.1	iPhone15,2	Pillow	730	5	0	in_bed
24	Europe/Vienna	com.neybox.Pillow	17.6.1	iPhone15,2	Pillow	745	8	1	in_bed
48	Europe/Podgorica	com.neybox.Pillow	18.0.0	iPhone15,2	Pillow	751	8	22	in_bed

	HKTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value
0	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	8	0	in_bed
4	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.3.0	iPhone15,2	iPhone Аннушка	17.3	5	22	in_bed
5	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	7	23	in_bed
7	Europe/Vienna	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.4.1	iPhone15,2	iPhone Аннушка	17.4.1	8	23	in_bed
8	Europe/Podgorica	com.apple.health.285A5E5B-B5D6-4182-BC34-9A821...	17.1.2	iPhone15,2	iPhone Аннушка	17.1.2	6	23	in_bed

	HKTimeZone	bundleIdentifier	operatingSystemVersion	productType	sourceName	version	timeEnd	timeStart	value
16	Europe/Podgorica	com.apple.health.D3500757-3FFE-4417-8F0A-FAD96...	11.0.0	Watch6,15	Анна's Apple Watch	11.0	23	23	light
30	Europe/Vienna	com.apple.health.D3500757-3FFE-4417-8F0A-FAD96...	11.0.0	Watch6,15	Анна's Apple Watch	11.0	0	0	light
40	Europe/Vienna	com.apple.health.D3500757-3FFE-4417-8F0A-FAD96...	11.0.0	Watch6,15	Анна's Apple Watch	11.0	0	0	light
44	Europe/Vienna	com.apple.health.D3500757-3FFE-4417-8F0A-FAD96...	11.0.0	Watch6,15	Анна's Apple Watch	11.0	0	23	awake
57	Europe/Vienna	com.apple.health.D3500757-3FFE-4417-8F0A-FAD96...	11.0.0	Watch6,15	Анна's Apple Watch	11.0	23	23	light

Здесь каждый источник регистрирует время отхода ко сну и время пробуждения, а часы Watch также резюмируют качество сна.

	bundleIdentifier	operatingSystemVersion	sourceName	version	timeEnd	timeStart
35	com.ouraring.oura	17.6.1	Oura	2408150708	7	22
36	com.apple.Health	18.0.0	Health	18.0	7	22
45	com.ouraring.oura	18.0.0	Oura	2409241358	9	23
98	com.ouraring.oura	18.0.0	Oura	2409241358	7	22
178	com.ouraring.oura	18.0.0	Oura	2409241358	7	0

Здесь 2 источника: ‘Health’, ‘Oura’.

	bundleIdentifier	operatingSystemVersion	sourceName	version	timeEnd	timeStart
36	com.apple.Health	18.0.0	Health	18.0	7	22

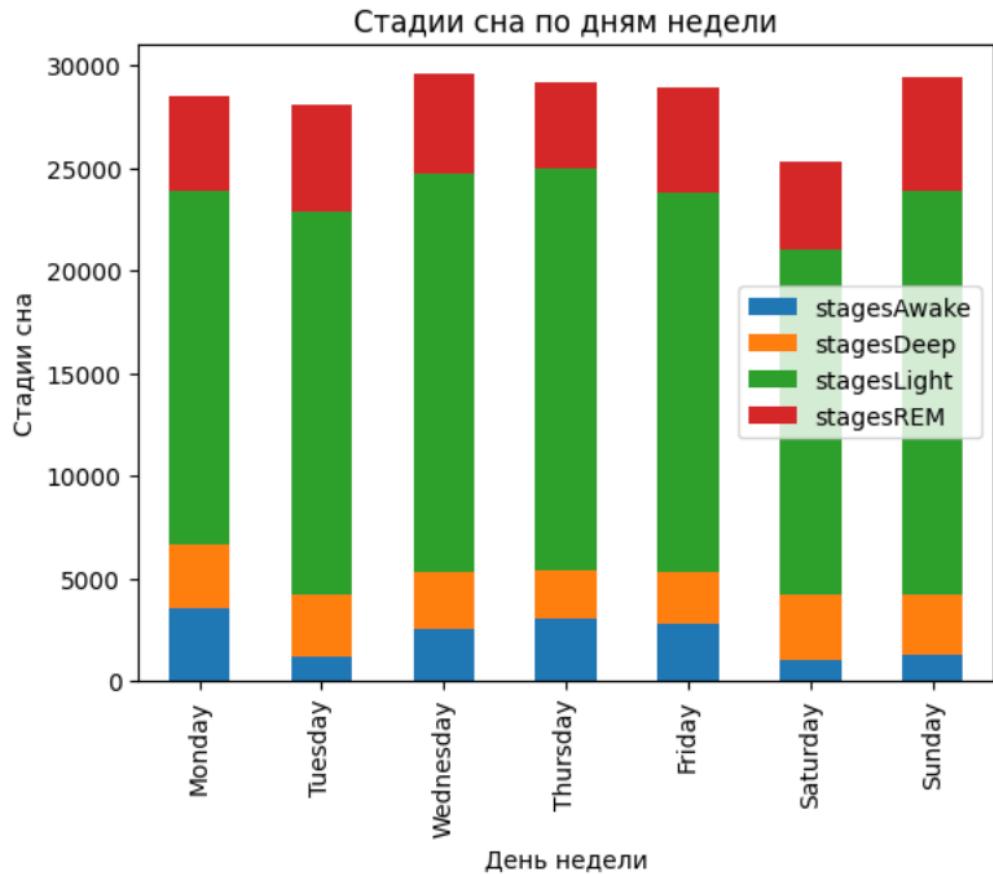
	bundleIdentifier	operatingSystemVersion	sourceName	version	timeEnd	timeStart
35	com.ouraring.oura	17.6.1	Oura	2408150708	7	22
45	com.ouraring.oura	18.0.0	Oura	2409241358	9	23
98	com.ouraring.oura	18.0.0	Oura	2409241358	7	22
178	com.ouraring.oura	18.0.0	Oura	2409241358	7	0
235	com.ouraring.oura	18.0.0	Oura	2409241358	6	23

Данные источники регистрируют только время отхода ко сну и время пробуждения.

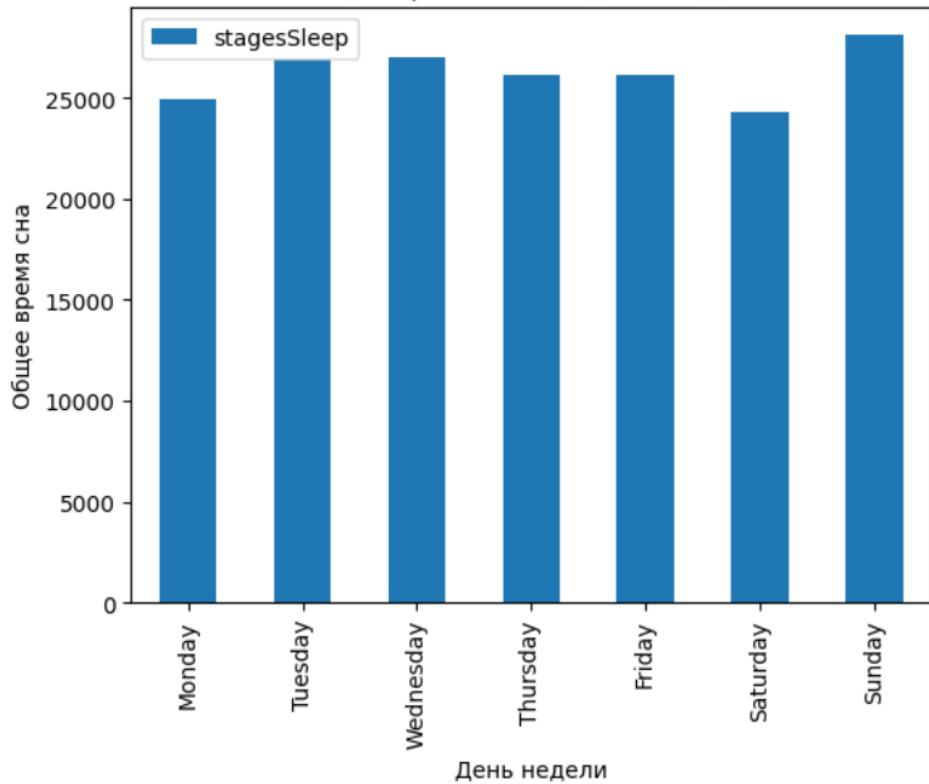
Выводы к разделу 2

После анализа результатов от разных источников, самый показательный оказался AutoSleep.

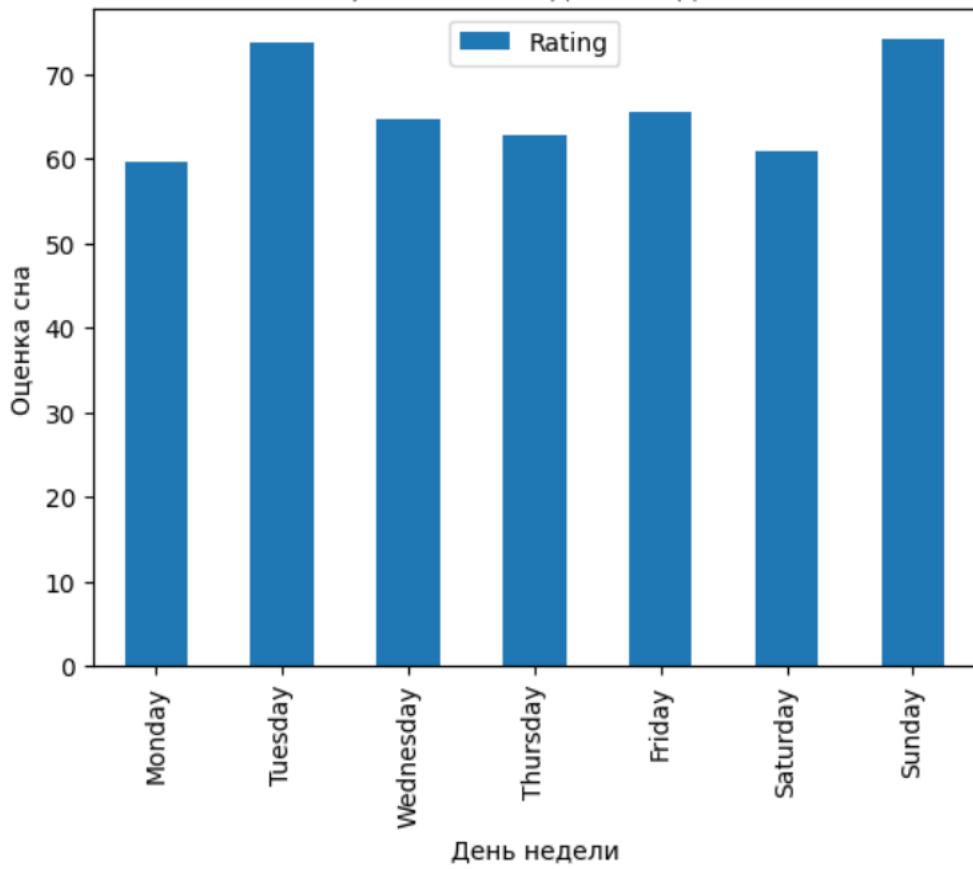
Раздел 3. Статистика



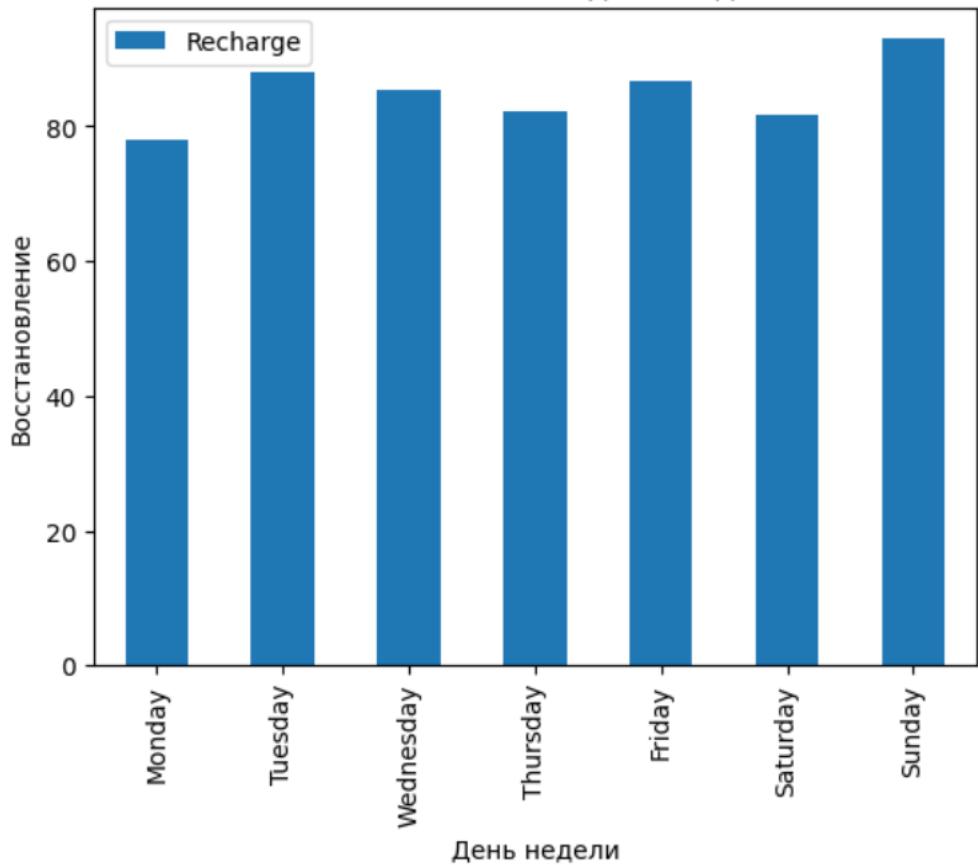
Общее время сна по дням недели



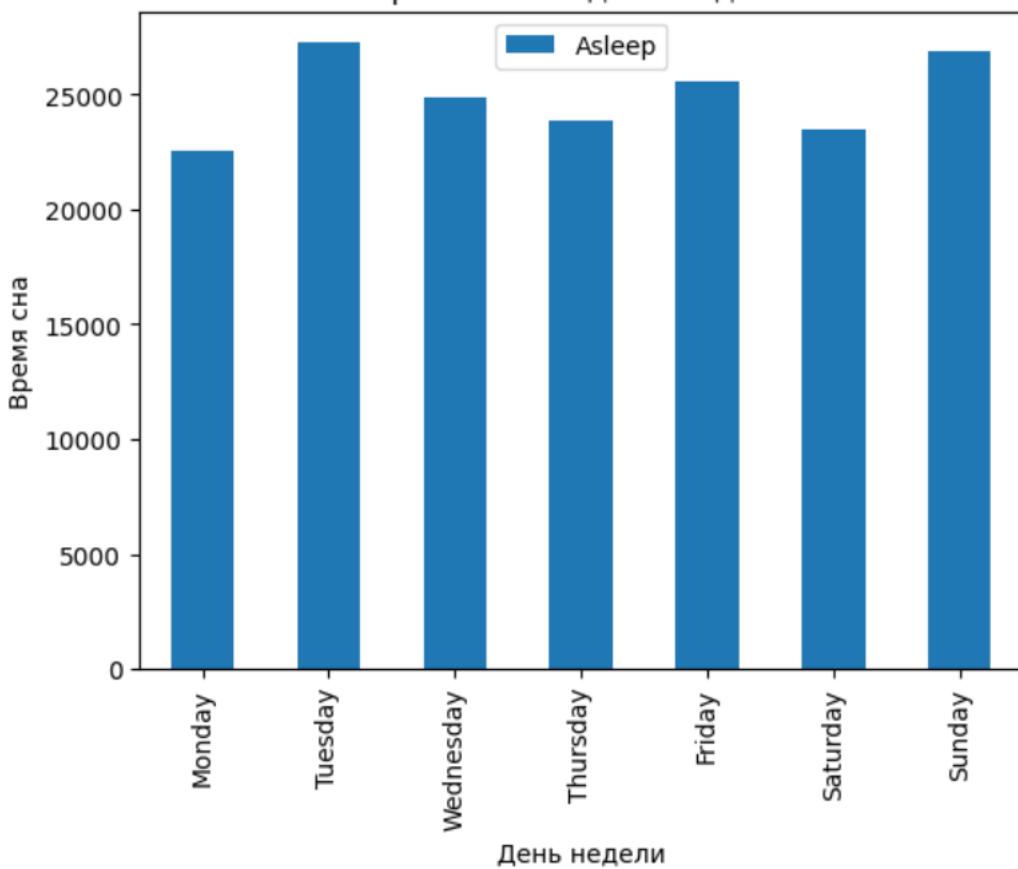
Оценка сна по дням недели



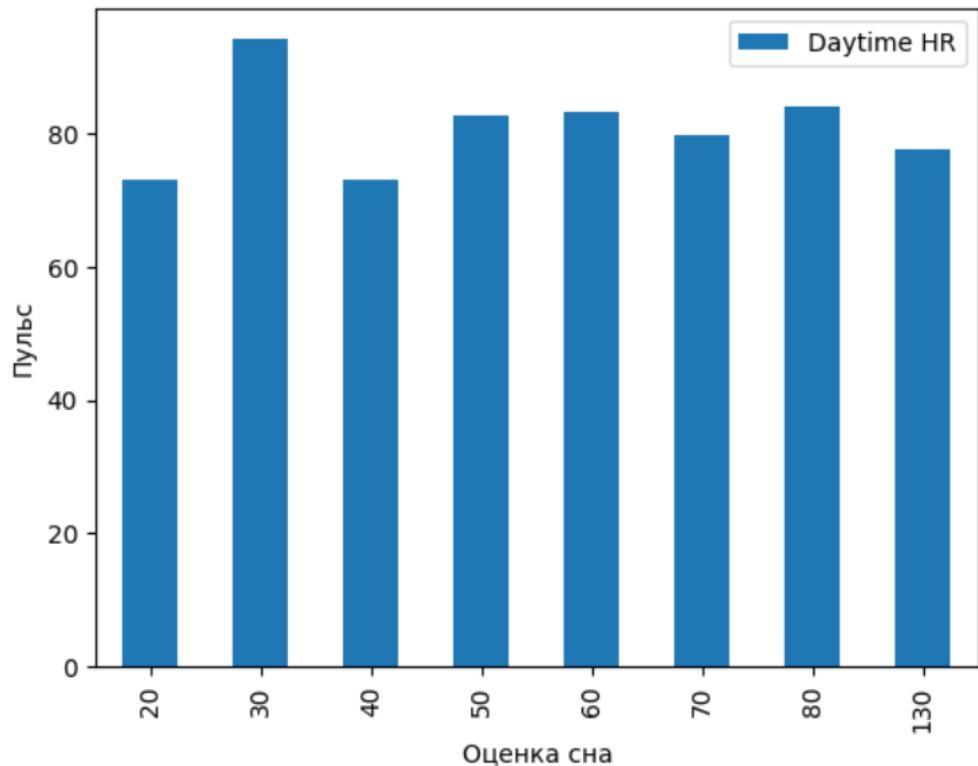
Восстановление по дням недели

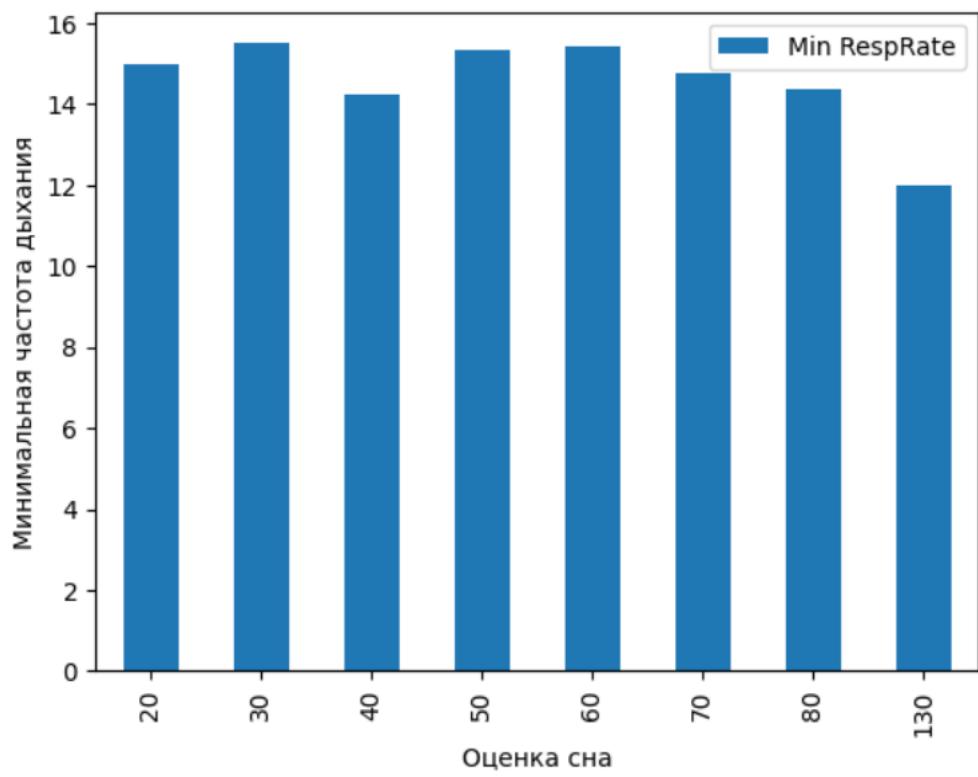
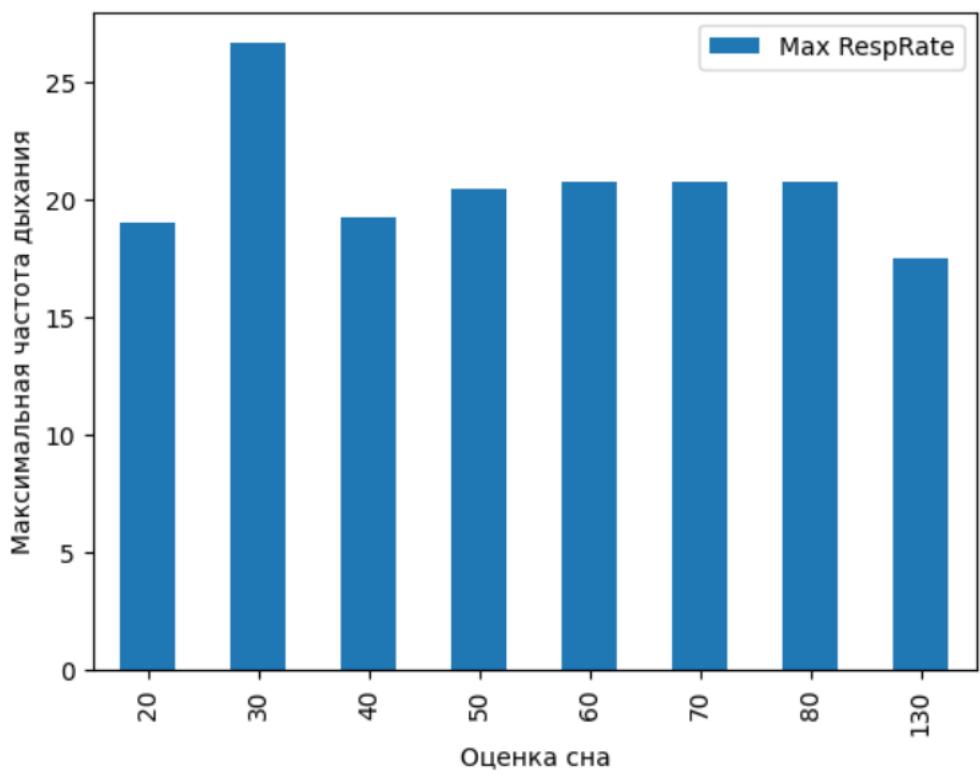


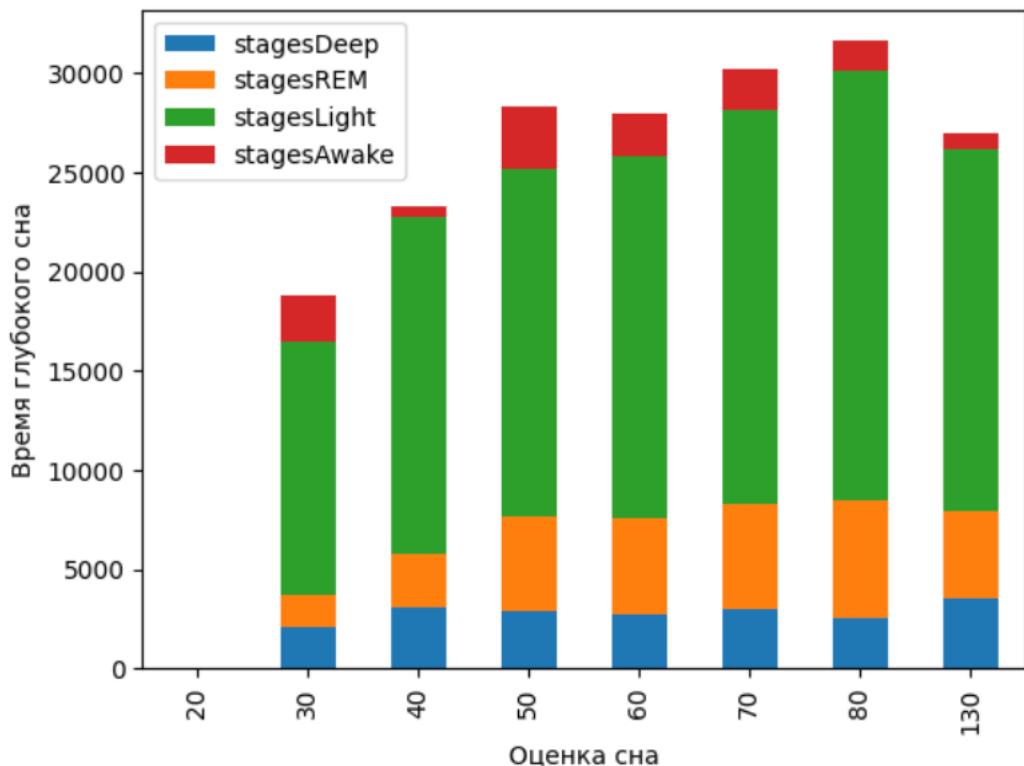
Время сна по дням недели



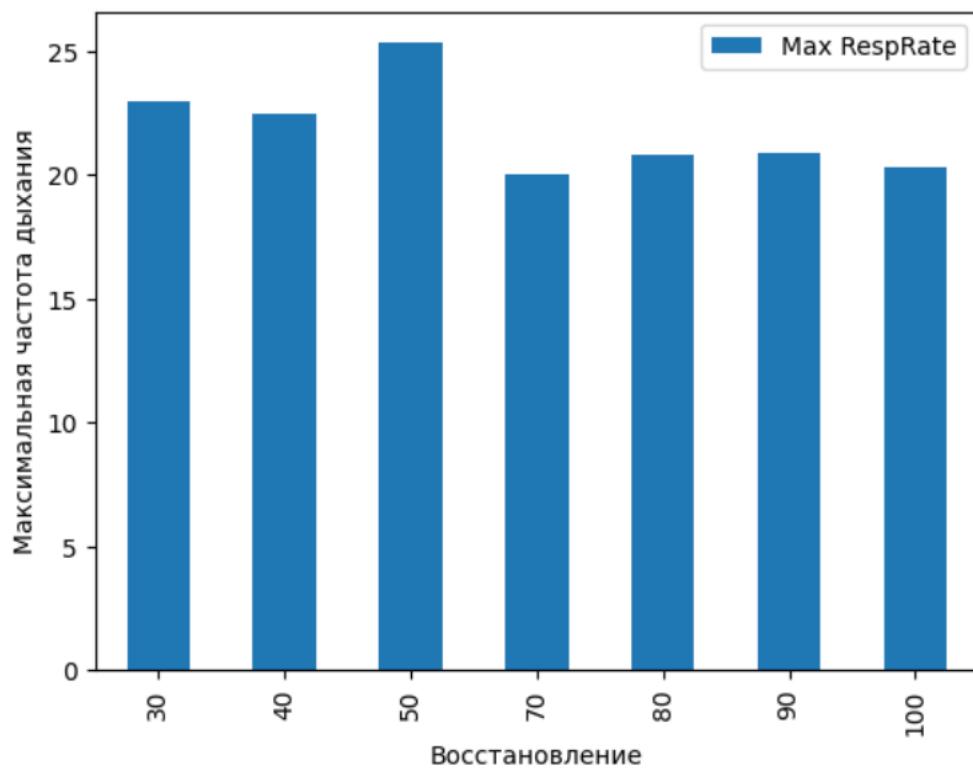
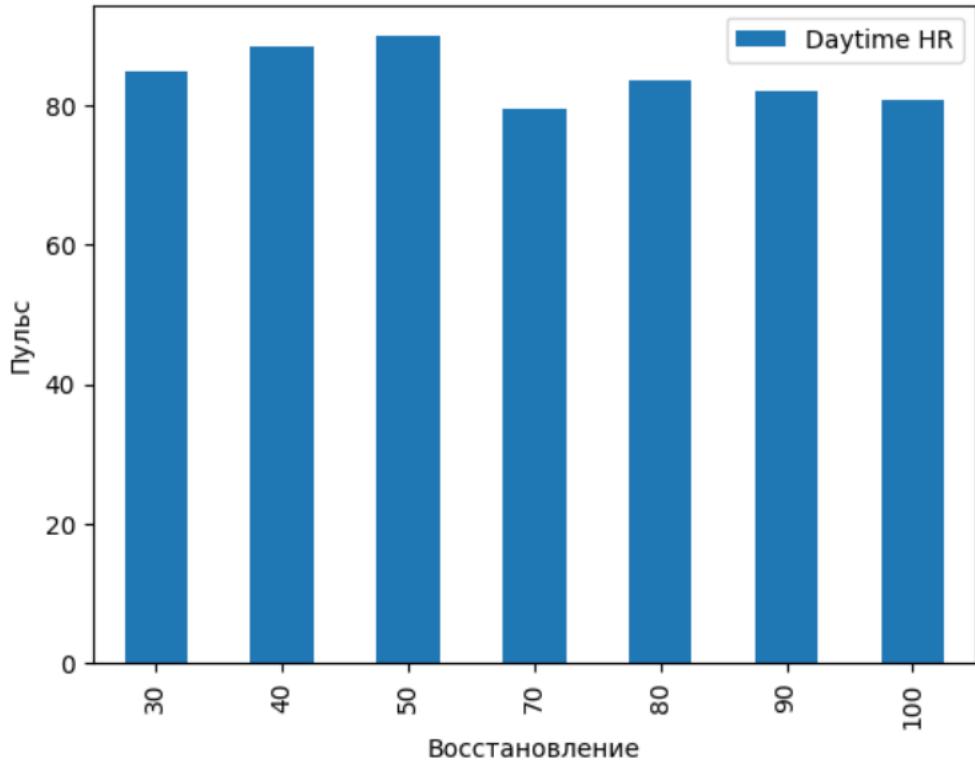
Судя по данным диаграммам, больше всего люди высыпаются в ночь со вторника на среду и с воскресенья на понедельник, а с субботы на воскресенье меньше всего. Стадии глубокого, легкого и быстрого сна примерно пропорциональны друг другу. Больший вклад в восстановление приносит глубокий сон.

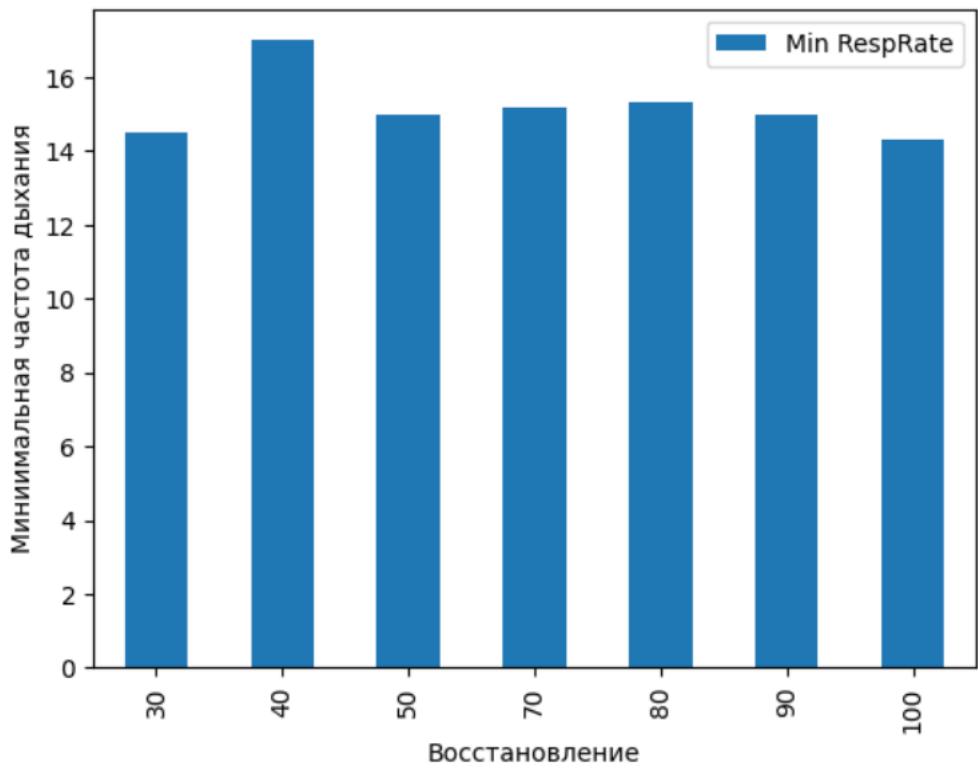




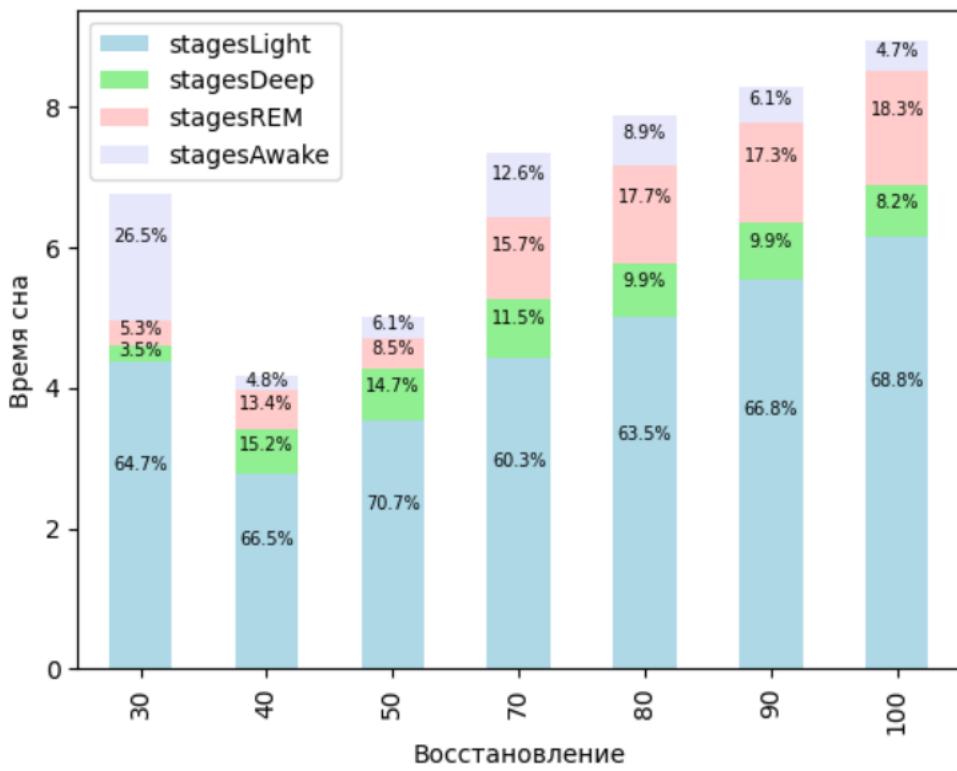


При показателе оценки сна в 30-40% пульс днем и максимальная частота дыхания во время сна повышенны, в отличие от более низкого показателя, что объясняется тем, что отсутствие сна получило оценку в 20-30%. При чрезмерном сне частота дыхания (как максимальная, так и минимальная) снижены.





Судя по диаграмме, до 50-60 % восстановления пульс довольно высокий, при этом возрастая, далее становится ниже. Частота дыхания до 50-60 % восстановления выше, чем после, где оно достаточно ровное.



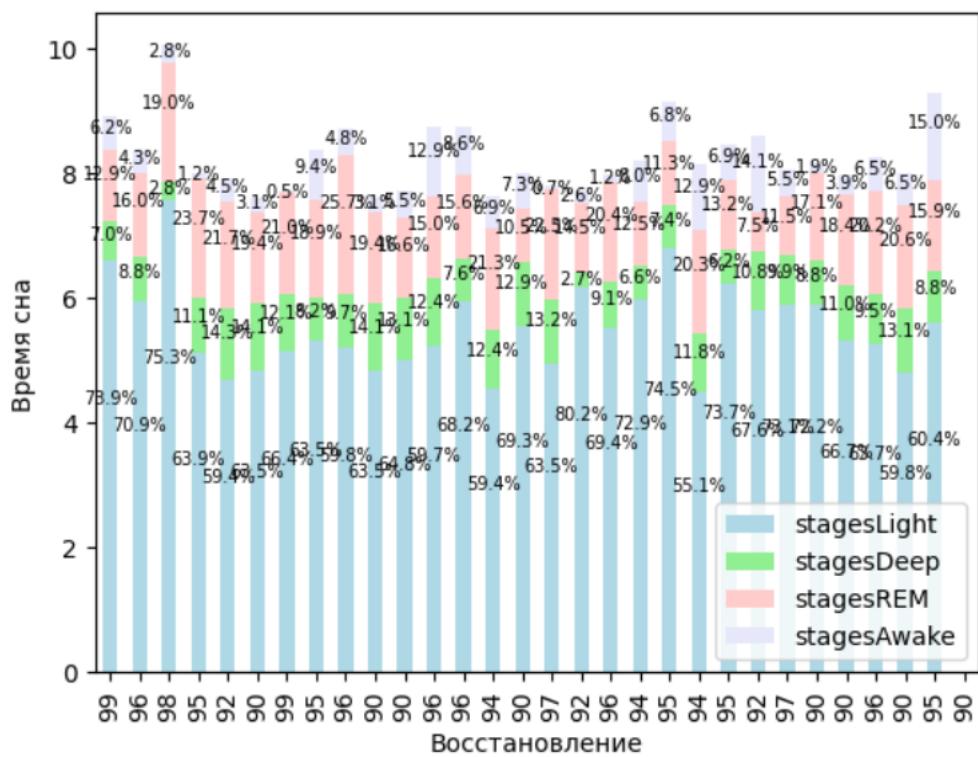
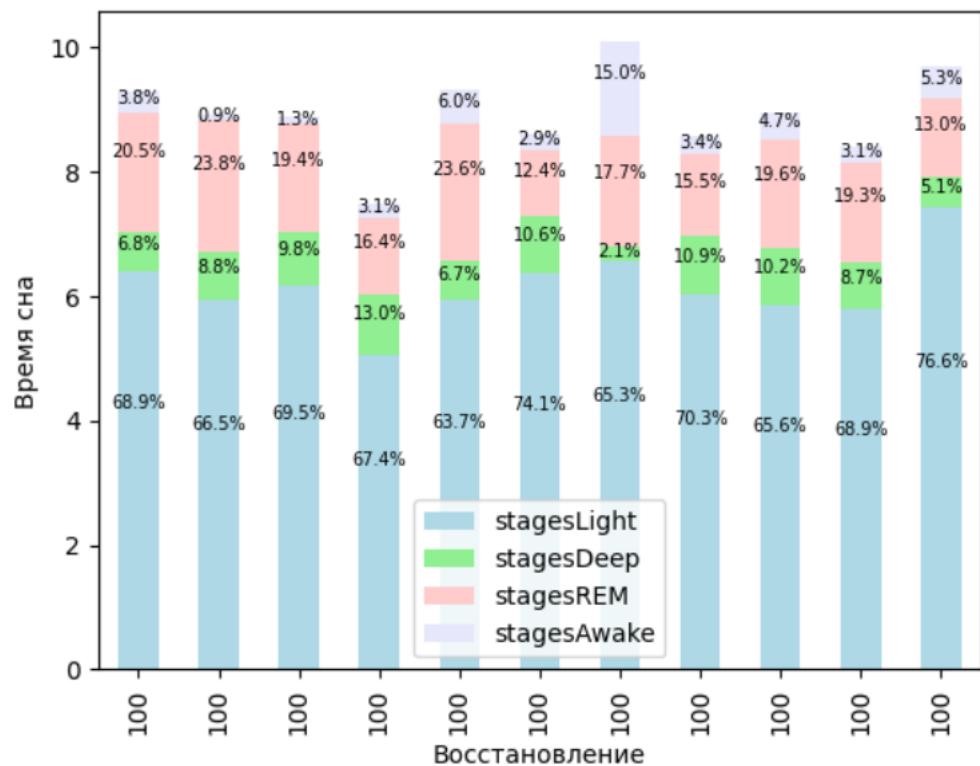
Согласно исследованиям фазы сна делятся в данном соотношении:

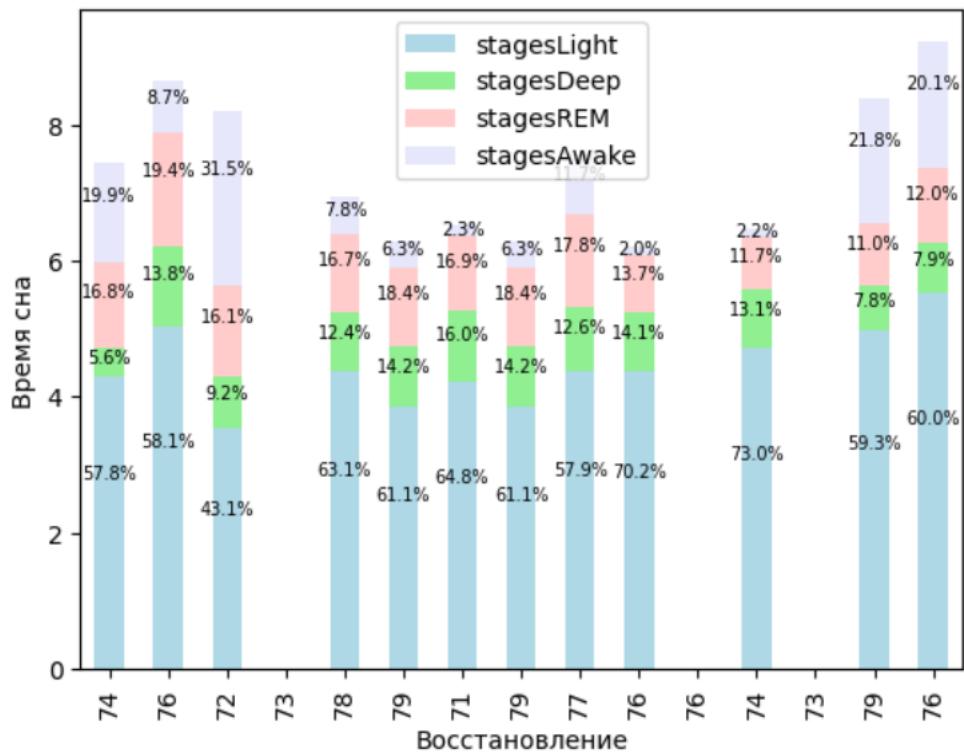
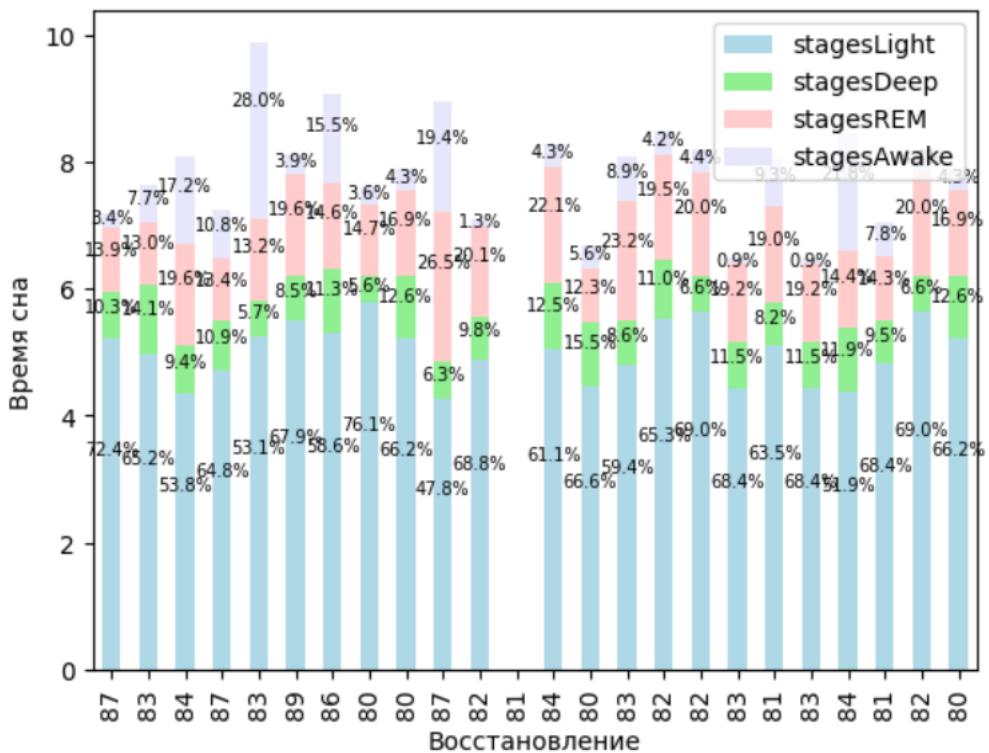
- Дремота (Non-REM1) и бодрствование 2-5%
- Легкий сон (Non-REM2) 45-60%
- Глубокий сон (Non-REM3) 15-25%
- Быстрый сон (REM) 20-25%

На графике средних значений времени сна показатели немного отличаются от данными.

Количество времени, проведенного в фазах глубокого и быстрого сна напрямую влияют на процесс восстановления. Глубокий сон обеспечивает физическое восстановление. В этот период происходит активная регенерация клеток, укрепление иммунной системы и восстановление энергии. Быстрый сон отвечает за эмоциональное состояние и когнитивные функции. Именно в этот период мозг обрабатывает информацию и упорядочивает её, что помогает укреплять память и учиться новому.

Посмотрим на показатели процессов сна, которые обеспечили 100% восстановление.





Итого:

- Сон с 100-процентным восстановлением либо 8+ часов, либо с хорошим соотношением глубокого сна и быстрого (3 к 2);

- У сна с восстановлением 90-99% примерно такая же ситуация соотношений глубокого и быстрого снов, но немного меньше общее время сна и легкий сон;
- Сон с восстановлением 80-89% имеет продолжительность сна от 6 до 8 часов. В остальном - так же, как предыдущий пункт, и точно так же меньше продолжительность легкого сна;
- Сон с восстановлением 70-79% имеет продолжительность от 5.8 до 7.8 часов. Еще короче легкий сон (в соотношении), во многих случаях достаточно большой процент стадий бодрствования.

Выводы к разделу 3

Все стадии сна влияют на восстановление организма. Наличие глубокого и быстрого сна хорошей пропорции в сравнении с общим временем сна, а также полный 8-часовой сон помогают восстановиться лучше. Недостаток сна повышает пульс на протяжении дня, а переизбыточный сон влияет на частоту дыхания. Больше всего люди высыпаются со вторника на среду и с воскресенья на понедельник.