Анализ данных интеллектуальных счетчиков Лондона

Чтобы лучше следить за потреблением энергии, правительство Великобритании хочет, чтобы поставщики энергии устанавливали интеллектуальные счетчики в каждом доме в Англии, Уэльсе и Шотландии. В перспективе для поставщиков энергии есть 26 миллионов домов, и в планах к 2020 году каждый дом должен иметь умный счетчик.

Вашему вниманию представлено исследование, основанное на данных, содержащих показания, снятые с интеллектуальных счетчиков, по энергопотреблению для 5566 лондонских домохозяйств, которые приняли участие в проекте **Low Carbon London**, возглавляемом британскими энергосистемами, в период с ноября 2011 года по февраль 2014 года.

В проекте приняли участие:

Project lider:

Фурман Юрий Анатольевич, yuri063@yandex.ru

Разработчики:

Шапилов Александр Викторович, sankas@lenta.ru

Яров Евгений Андреевич, chatobec@yandex.ru

Зинович Эдуард Владимирович, bozon5@yandex.ru

Technical writer

Боткин Александр Сергеевич, spratlist@yandex.ru

Данное исследование проведено в соответствии с представленным **ниже планом**:

1. Первичный анализ имеющихся данных:

1.1. Выборка основных наборов данных для дальнейшего рассмотрения:

- informations_households.csv;
- acorn_details.csv;
- daily_dataset.csv;
- uk_bank_holidays.csv;
- weather_daily_darksky.csv;
- weather_hourly_darksky.csv;
- halfhourly_dataset;
- hhblock_dataset/block_xxx.

2. Подготовка и анализ данных для исследования:

2.1. Анализ распределения домохозяйств по группам;

2.2. Анализ полноты представленных данных об энергопотреблении;

2.3. Определение временного интервала (окна), достаточного для построения выборки наблюдений;

2.3. Финальное редактирование полученных данных

3. Анализ энергопотребления:

3.1. Произведем анализ зависимости суммарного потребления энергии от среднесуточной температуры. Найдем средний прирост потребления энергии при изменении температуры.

3.1.1. Построим график зависимости суммарного потребления энергии и среднесуточной температуры от времени.

3.1.2. Рассчитаем средний прирост потребления энергии при изменении среднедневной температуры на 1 градус Цельсия.

3.2. Сделаем сравнение потребления энергии в выходные/праздничные и рабочие дни

3.3. Для каждой из пяти групп потребителей проведение анализа суточного энергопотребления в течение 2013 года;

3.3.1. Сравнение ночного, дневного и суммарного энергопотребления

3.3.2. Сравнение энергопотребления по периодам T1, T2 и T3

3.4. Для каждой из пяти групп потребителей построение графиков зависимости суточного потребления энергии от длины светового дня:

3.4.1. Построение графиков зависимости суточного потребления энергии от длины светового дня, и от среднесуточной температуры

3.4.2. Расчёт среднего прироста потребления энергии при изменении среднедневной температуры комфорта и длительности светового дня

3.5. Исследование зависимости суммарного потребления энергии в течение года от времени суток; от дня недели и месяца года:

3.5.1. Тепловая карта потребления электроэнергии - время года/время суток

3.5.2. Тепловая карта потребления электроэнергии - дни недели/недели года

3.6. Подробное исследование зависимости потребления энергии от среднесуточных температур в течение года:

3.6.1. Построение регрессионной модели зависимости энергопотребления от температуры наружного воздуха

3.6.2. Нахождение оценки R-квадрат (коэффициент детерминации; в Python - r2_score)

3.6.3. Нахождение доли домовладений, предположительно использующих электроэнергию для отопления.

4. Итоговые выводы.

1. Обзор представленных данных

Набор данных, который мы использовали в проекте, можно найти по ссылке:

https://www.kaggle.com/jeanmidev/smart-meters-in-london

Из предоставленного набора для нашего исследования мы выбрали данные представленные в следующих файлах:

1. *informations_households.csv*

Файл содержит *информацию о датчиках энергопотребления* и краткие сведения о домах, в которых эти датчики установлены.

	LCLid	stdorToU	Acorn	Acorn_grouped	file
0	MAC005492	ToU	ACORN-	ACORN-	block_0
1	MAC001074	ToU	ACORN-	ACORN-	block_0
2	MAC000002	Std	ACORN-A	Affluent	block_0
3	MAC003613	Std	ACORN-A	Affluent	block_0
4	MAC003597	Std	ACORN-A	Affluent	block_0

Столбцы датафрейма:

- **LCLid –** id датчика;
- **stdorToU –** форма оплаты за электроэнергию в доме, в котором установлен датчик (**Std** – стандартная, **ToU** – оплата зависит от времени суток);
- **Acorn, Acorn_grouped** – информация о том, к какой категории потребителей по системе *ACORN* относится семья, проживающая в данном доме;
- **file** – имя файла, содержащего показатели счетчиков.

2. *acorn_details.csv*

Файл содержит данные о группах потребителей согласно классификации **ACORN**. При характеристике групп в данном датафрейме используется сравнение каждой группы с общенациональными показателями. Так, если значение ячейки по какому-либо показателю составляет 150, это означает, что в рассматриваемой группе этот показатель встречается в 1,5 раза чаще, чем в целом по стране.

	MAIN CATEGORIES	CATEGORIES	REFERENCE	ACORN-A	ACORN-B	ACORN-C	ACORN-D	ACORN-E	ACORN-F	ACORN-G	ACORN-H	ACORN-I	ACORN-J	ACORN-K
0	POPULATION	Age	Age 0-4	77.0	83.0	72.0	100.0	120.0	77.0	97.0	97.0	63.0	119.0	67.0
1	POPULATION	Age	Age 5-17	117.0	109.0	87.0	69.0	94.0	95.0	102.0	106.0	67.0	95.0	64.0
2	POPULATION	Age	Age 18-24	64.0	73.0	67.0	107.0	100.0	71.0	83.0	89.0	62.0	104.0	459.0
3	POPULATION	Age	Age 25-34	52.0	63.0	62.0	197.0	151.0	66.0	90.0	88.0	63.0	132.0	145.0
4	POPULATION	Age	Age 35-49	102.0	105.0	91.0	124.0	118.0	93.0	102.0	103.0	76.0	111.0	67.0

Столбцы датафрейма:

- **MAIN CATEGORIES, CATEGORIES, REFERENCE** – показатели, по которым осуществляется сравнение групп;
- **ACORN-A, ACORN-B, ACORN-C, ACORN-D, ACORN-E, ACORN-F, ACORN-G, ACORN-H, ACORN-I, ACORN-J, ACORN-K, ACORN-L, ACORN-M, ACORN-N, ACORN-O, ACORN-P, ACORN-Q** – частота встречаемости каждого из рассмотренных признаков в каждой группе.

3. *daily_dataset.csv*

Файл содержит *обобщенные данные об энергопотреблении* за каждые сутки с датчиков, установленных в домах.

	LCLid	day	energy_median	energy_mean	energy_max	energy_count	energy_std	energy_sum	energy_min
0	MAC000131	2011-12-15	0.4850	0.432045	0.868	22	0.239146	9.505	0.072
1	MAC000131	2011-12-16	0.1415	0.296167	1.116	48	0.281471	14.216	0.031
2	MAC000131	2011-12-17	0.1015	0.189812	0.685	48	0.188405	9.111	0.064
3	MAC000131	2011-12-18	0.1140	0.218979	0.676	48	0.202919	10.511	0.065
4	MAC000131	2011-12-19	0.1910	0.325979	0.788	48	0.259205	15.647	0.066

Столбцы датафрейма:

- **LCLid** – id датчика;
- **day** – дата;
- **energy_median** – медиана суточных показателей;
- **energy_mean** – среднее арифметическое суточных показателей;
- **energy_max** – максимальное значение энергопотребления, зафиксированное датчиком в течение дня;
- **energy_count** – количество показателей, зафиксированных датчиком в течение дня;
- **energy_std** – стандартное отклонение;
- **energy_sum** – сумма значений всех показателей;
- **energy_min** – минимальное значение энергопотребления, зафиксированное датчиком в течение дня.

4. *uk_bank_holidays.csv*

Файл содержит *информацию о праздничных днях* на период наблюдения.

	Bank holidays	Type
0	2012-12-26	Boxing Day
1	2012-12-25	Christmas Day
2	2012-08-27	Summer bank holiday
3	2012-05-06	Queen?s Diamond Jubilee (extra bank holiday)
4	2012-04-06	Spring bank holiday (substitute day)

Столбцы датафрейма:

- **Bank_holidays** – дата выходного/праздничного дня;
- **Type** – название праздничного дня.

5. *weather_daily_darksky.csv*

Файл содержит *обобщённые данные о погоде за день*.

	0	1	2
temperatureMax	11.96	8.59	10.33
temperatureMaxTime	2011-11-11 23:00:00	2011-12-11 14:00:00	2011-12-27 02:00:00
windBearing	123	198	225
icon	fog	partly-cloudy-day	partly-cloudy-day
dewPoint	9.4	4.49	5.47
temperatureMinTime	2011-11-11 07:00:00	2011-12-11 01:00:00	2011-12-27 23:00:00
cloudCover	0.79	0.56	0.85
windSpeed	3.88	3.94	3.54
pressure	1016.08	1007.71	1032.76
apparentTemperatureMinTime	2011-11-11 07:00:00	2011-12-11 02:00:00	2011-12-27 22:00:00
apparentTemperatureHigh	10.87	5.62	10.33
precipType	rain	rain	rain
visibility	3.3	12.09	13.39
humidity	0.95	0.88	0.74
apparentTemperatureHighTime	2011-11-11 19:00:00	2011-12-11 19:00:00	2011-12-27 14:00:00
apparentTemperatureLow	10.87	-0.64	5.52
apparentTemperatureMax	11.96	5.72	10.33
uvIndex	1	1	0
time	2011-11-11 00:00:00	2011-12-11 00:00:00	2011-12-27 00:00:00
sunsetTime	2011-11-11 16:19:21	2011-12-11 15:52:53	2011-12-27 15:57:56
temperatureLow	10.87	3.09	8.03
temperatureMin	8.85	2.48	8.03
temperatureHigh	10.87	8.59	10.33
sunriseTime	2011-11-11 07:12:14	2011-12-11 07:57:02	2011-12-27 08:07:06
temperatureHighTime	2011-11-11 19:00:00	2011-12-11 14:00:00	2011-12-27 14:00:00
uvIndexTime	2011-11-11 11:00:00	2011-12-11 12:00:00	2011-12-27 00:00:00
summary	Foggy until afternoon.	Partly cloudy throughout the day.	Mostly cloudy throughout the day.
temperatureLowTime	2011-11-11 19:00:00	2011-12-12 07:00:00	2011-12-27 23:00:00
apparentTemperatureMin	6.48	0.11	5.59
apparentTemperatureMaxTime	2011-11-11 23:00:00	2011-12-11 20:00:00	2011-12-27 02:00:00
apparentTemperatureLowTime	2011-11-11 19:00:00	2011-12-12 08:00:00	2011-12-28 00:00:00
moonPhase	0.52	0.53	0.1

Столбцы датафрейма:

- **temperatureMax** – максимальное значение температуры воздуха;
- **temperatureMaxTime** – время, когда была зафиксирована максимальная температура воздуха;
- **windBearing** – направление ветра (по азимуту);
- **icon** – стандартизованное словесное описание погодных условий;
- **dewPoint** – точка росы;
- **temperatureMinTime** – время, когда была зафиксирована минимальная температура воздуха;
- **cloudCover** – облачность;
- **windSpeed** – скорость ветра;
- **pressure** – атмосферное давление;
- **apparentTemperatureMinTime** – время, когда была зафиксирована минимальная температура комфорта;

- **apparentTemperatureHigh** – дневная температура комфорта;
- **precipType** – тип осадков;
- **visibility** – видимость (в милях);
- **humidity** – относительная влажность;
- **apparentTemperatureHighTime** – время, когда была зафиксирована дневная температура комфорта;
- **apparentTemperatureLow** – ночная температура комфорта;
- **apparentTemperatureMax** – максимальная температура комфорта;
- **uvIndex** – UV-индекс;
- **time** – время начала сбора данных;
- **sunsetTime** – время заката;
- **temperatureLow** – минимальная ночная температура;
- **temperatureMin** – минимальная температура за сутки;
- **temperatureHigh** – максимальная дневная температура;
- **sunriseTime** – время восхода;
- **temperatureHighTime** – время, когда была зафиксирована максимальная дневная температура;
- **uvIndexTime** – время, когда был зафиксирован максимальный UV-индекс;
- **summary** – словесное описание погоды в течение дня (не рекомендуется использовать для автоматизированного анализа!);
- **temperatureLowTime** – время, когда была зафиксирована минимальная ночная температура;
- **apparentTemperatureMin** – минимальная температура комфорта;
- **apparentTemperatureMaxTime** – время, когда была зафиксирована максимальная температура комфорта за сутки;
- **apparentTemperatureLowTime** – время, когда была зафиксирована минимальная ночная температура комфорта;
- **moonPhase** – фаза луны.

6. *weather_hourly_darksky.csv*

Файл содержит *почасовые сведения о погоде*.

	visibility	windBearing	temperature	time	dewPoint	pressure	apparentTemperature	windSpeed	precipType	icon	humidity	summary
0	5.97	104	10.24	2011-11-11 00:00:00	8.86	1016.76	10.24	2.77	rain	partly-cloudy-night	0.91	Partly Cloudy
1	4.88	99	9.76	2011-11-11 01:00:00	8.83	1016.63	8.24	2.95	rain	partly-cloudy-night	0.94	Partly Cloudy
2	3.70	98	9.46	2011-11-11 02:00:00	8.79	1016.36	7.76	3.17	rain	partly-cloudy-night	0.96	Partly Cloudy
3	3.12	99	9.23	2011-11-11 03:00:00	8.63	1016.28	7.44	3.25	rain	fog	0.96	Foggy
4	1.85	111	9.26	2011-11-11 04:00:00	9.21	1015.98	7.24	3.70	rain	fog	1.00	Foggy

Столбцы датафрейма:

- **visibility** – видимость в милях;
- **windBearing** – направление ветра (по азимуту);
- **temperature** – температура воздуха;
- **time** – время записи показателей;
- **dewPoint** – точка росы;
- **pressure** – атмосферное давление;
- **apparentTemperature** – температура комфорта;
- **windSpeed** – скорость ветра;
- **precipType** – тип осадков;
- **icon** – стандартизованное словесное описание погодных условий;

- **humidity** – относительная влажность;
- **summary** – не стандартизованное словесное описание погодных условий.

7. *halfhourly_dataset*

Папка включает 112 файлов, содержащих данные об энергопотреблении, получаемые с каждого счетчика 1 раз в 30 минут.

	LCLid	tstp	energy(kWh/hh)
0	MAC000002	2012-10-12 00:30:00.0000000	0
1	MAC000002	2012-10-12 01:00:00.0000000	0
2	MAC000002	2012-10-12 01:30:00.0000000	0
3	MAC000002	2012-10-12 02:00:00.0000000	0
4	MAC000002	2012-10-12 02:30:00.0000000	0

Столбцы датафрейма:

- **LCLid** – id датчика;
- **tstp** – дата и время фиксации показателей;
- **energy**(kWh/hh) – уровень энергопотребления.

8. *hhblock_dataset*

Папка включает 112 файлов с данными о получасовом энергопотреблении одного домохозяйства в день, например, столбец hh_0 является потреблением между 00:00 и 00:30.

	LCLid	day	hh_0	hh_1	hh_2	hh_3	hh_4	hh_5	hh_6	hh_7	...	hh_38	hh_39	hh_40	hh_41	hh_42	hh_43	hh_44	hh_45	hh_46	hh_47
0	MAC000002	2012-10-13	0.263	0.269	0.275	0.256	0.211	0.136	0.161	0.119	...	0.918	0.278	0.267	0.239	0.230	0.233	0.235	0.188	0.259	0.250
1	MAC000002	2012-10-14	0.262	0.166	0.226	0.088	0.126	0.082	0.123	0.083	...	1.075	0.956	0.821	0.745	0.712	0.511	0.231	0.210	0.278	0.159
2	MAC000002	2012-10-15	0.192	0.097	0.141	0.083	0.132	0.070	0.130	0.074	...	1.164	0.249	0.225	0.258	0.260	0.334	0.299	0.236	0.241	0.237
3	MAC000002	2012-10-16	0.237	0.237	0.193	0.118	0.098	0.107	0.094	0.109	...	0.966	0.172	0.192	0.228	0.203	0.211	0.188	0.213	0.157	0.202
4	MAC000002	2012-10-17	0.157	0.211	0.155	0.169	0.101	0.117	0.084	0.118	...	0.223	0.075	0.230	0.208	0.265	0.377	0.327	0.277	0.288	0.256

Столбцы датафрейма:

- **LCLid** – id датчика;
- **day** – дата;
- **hh_0** … **hh_47** – столбцы с получасовыми данными энергопотребления.

Примечание:

Т.к. наборы данных в папках ***halfhourly_dataset*** и ***hhblock_dataset*** идентичны (с точностью до транспонирования), то для дальнейшего рассмотрения оставим один из них - ***hhblock_dataset***

2. Подготовка и анализ данных для исследования

В данной части проекта оценим полноту и корректность представленных данных, опишем логику, которой руководствовалась команда, и приведём данные в соответствие с этой логикой.

2.1. Анализ распределения домохозяйств по группам

Из описания датасета известно, что все 5566 домохозяйств в наборе ***hhblock_dataset/block_xxx*** разбиты на 112 блоков (файлов), примерно, по 50 в каждом, по следующему принципу:

- сбор всех данных от конкретного домохозяйства в одном блоке (файле)
- в одном блоке желательно одна группа потребителей по финансово-социальному статусу (Acorn).

Принадлежность домохозяйств к группам потребления описана в наборе ***informations_households.csv***.

Найдем соответствие между *группами* **Acorn_grouped** и *блоками вида* ***block_xxx,*** т.е. сколько всего *групп* **Acorn_grouped** и каким *блокам* они соответствуют*:*

Acorn_grouped file	ACORN-	ACORN-U	Adversity	Affluent	Comfortable
block_0	2	0	0	48	0
block_1	0	0	0	50	0
block_10	0	0	0	50	0
block_100	0	0	50	0	0
block_101	0	0	50	0	0

При внимательном рассмотрении мы могли видеть, что в **informations_households.csv** есть неточности в описании принадлежности к **Acorn_grouped** и сделано довольно *грубое объединение по группам потребления* (объединение нескольких **Acorn** в одну группу).

Исправим это. Сделаем перегруппировку в соответствии с **Acorn-User-guide** и посмотрим, как распределились домовладения по новым группам в соответствии с их принадлежностью к группе **Acorn**:

- **Affluent:** 'acorn-a', 'acorn-b', 'acorn-c' *(богатые люди);*
- **Rising:** 'acorn-d', 'acorn-e' *(успешные, с растущим благосостоянием);*
- **Comfortable:** 'acorn-f', 'acorn-g', 'acorn-h', 'acorn-i', 'acorn-j' *(владельцы и жители комфортабельного жилья);*
- **Stretched:** 'acorn-k', 'acorn-l', 'acorn-m', 'acorn-n' *(финансово ограниченные);*
- **Adversity:** 'acorn-o', 'acorn-p', 'acorn-q' *(городская беднота);*
- **NP-Household:** все остальные *(общественные домовладения);*

Acorn_grouped	LCLid	stdorToU	Acorn	file
Adversity	1044	1044	1044	1044
Affluent	333	333	333	333
Comfortable	1507	1507	1507	1507
NP-Household	51	51	51	51
Rising	1859	1859	1859	1859
Stretched	772	772	772	772

Рассмотрим подробнее **Not Private Households** ("общественные домовладения"). Они включают три типа потребителей:

- активная группа (военные базы, отели, общежития, детские приюты, т.е. где жители постоянно меняются, но при этом ведут активный образ жизни и потребления),
- не активная группа (больницы, дома престарелых, тюрьмы и т.п.);
- бизнес учреждения (не жилье).

Т.к. эти домовладения не имеют постоянных жителей, не являются жильем в прямом смысле слова, и значительную часть времени могут простаивать, предлагается исключить их из рассмотрения, как не имеющие общих критериев оценки факторов потребления электроэнергии с остальными домовладениями.

Как видно из таблицы выше, доля домовладений типа **Not Private Households** в общем количестве домовладений не значительна (*51 домовладение*). Занесем в список ***list_hh_NP_less*** перечень домохозяйств из всех *Acorn_grouped*, кроме *NP-Household.*

2.2. Анализ полноты представленных данных об энергопотреблении

Прежде чем проводить анализ любых зависимостей энергопотребления от погодных, календарных, социальных факторов и привычек (владельцев домохозяйств) изучим вопрос о полноте представленных данных.

2.2.1. Посчитаем количество домохозяйств и дней, в которых имеются неполные измерения (меньше 48 измерений в сутки).

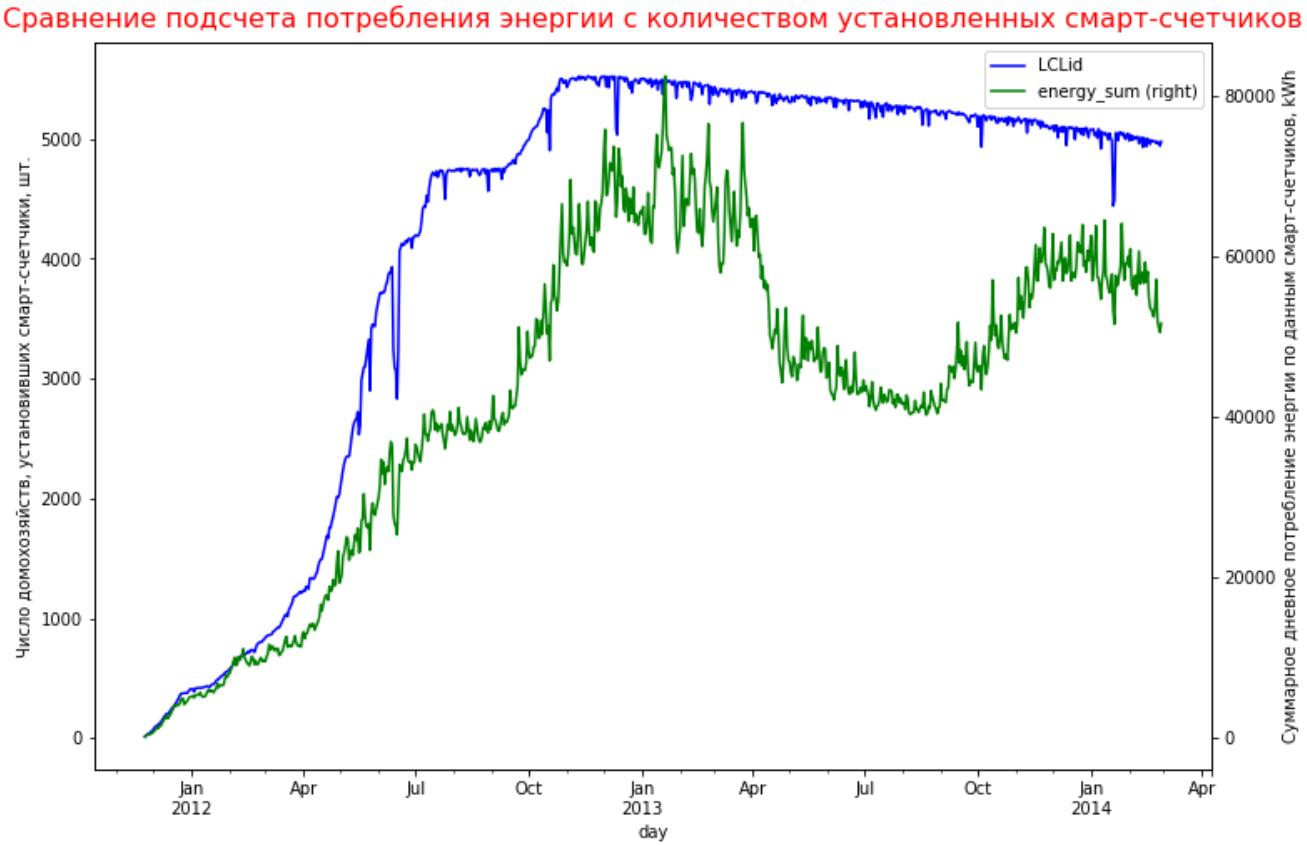
Установлено, что общее количество измерений, представленных в наборе данных **daily_dataset.csv** равняется **3510433**, а измерений с неполной информацией (меньше 48 измерений в сутки) равняется **41081**, что в процентном соотношении составляет 1.17%.

Т.к. доля неполных измерений чуть больше 1%, то в дальнейшем можно ими пренебречь с целью уменьшения влияния на погрешность суточных данных.

2.2.2. Рассмотрим на сколько равномерно и в каком объеме представлены (количественно) домовладения, проводившие измерения в течении всего исследуемого периода.

Построим на одном графике две кривые:

- *количество счетчиков* (с которых снимали показания за один день) в течение всего периода наблюдений (X: даты; Y - количество счетчиков в день);
- *суммарное потребление электроэнергии в день* в течение всего периода наблюдений (X: даты; Y - суммарное энергопотребление в день).



Вывод: *по построенному графику видно, что на начальном этапе потребление энергии зависит только от роста числа счетчиков. При этом, например, сезонный фактор не просматривается. Этот временной этап, очевидно, соответствует периоду установки счетчиков и наладке системы* (с 11.2011 до, примерно, 09.2012). *Данные за этот период надо также исключить из рассмотрения.*

2.2.3. Определим временной интервал (окно), который будем считать достаточным для построения выборки наблюдений

Статистически состоятельным (достаточным) можно считать период наблюдений, охватывающий все сезонные колебания погодных условий, и другие циклически повторяющиеся факторы, которые влияют на потребление электроэнергии. Логично, что за такой *период* можно принять *один календарный год*. Из всей выборки для этих целей лучше всего *подходят данные за 2013 год*, как наиболее полные, т.к. измерения по остальным годам либо представлены не полно *(11.2011 - 09.2012)*, либо не всеми сезонами и месяцами *(2011, 2014)*.

В результате *сужения* периода наблюдений (окна наблюдений) **до** одного **2013 года** (при условии полноты представленных измерений), количество домохозяйств в наборе данных *уменьшилось* **до 5528** (т.е. мы *"потеряли"* **38 домохозяйств**).

2.2.4. Исключим из рассмотрения домохозяйства, для которых количество дней наблюдений (сбора данных) в 2013 году существенно меньше, чем количество дней в году

Предположим, что достаточным периодом наблюдений является количество рабочих и выходных дней в году (за исключением праздничных выходных дней).

Используя данные из файла **uk_bank_holidays.csv** мы получили **8** *праздничных дней* в 2013 году, и, соответственно, достаточный период наблюдений составит **357 дней**.

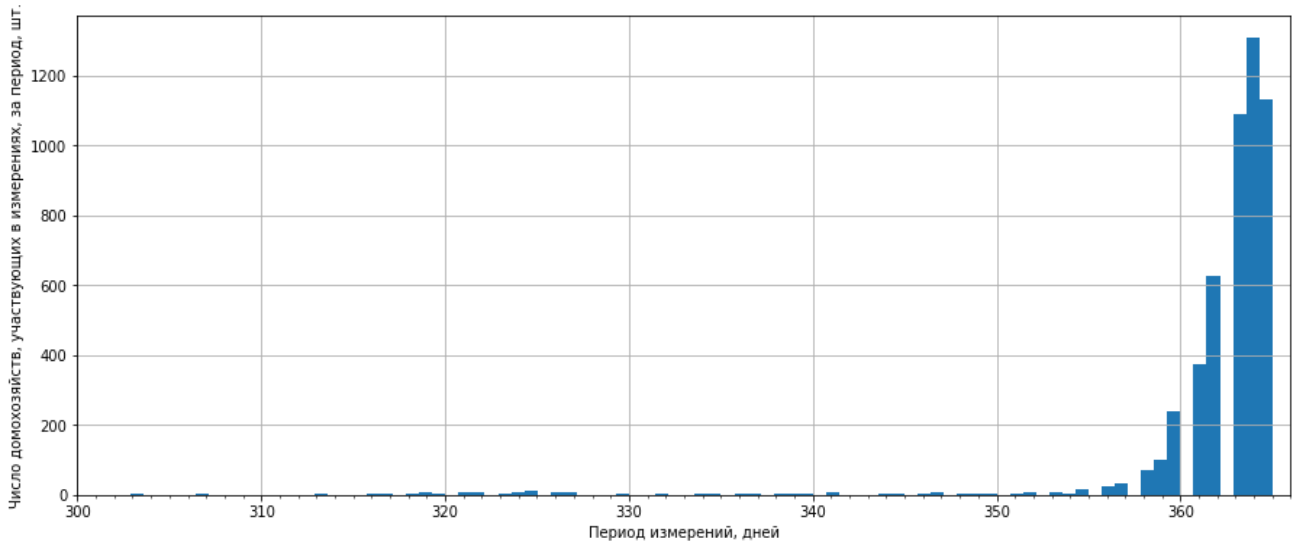
Проверим гипотезу о влиянии праздничных дней на период наблюдения. Ранжируем (упорядочим) все дни года, по увеличению количества домохозяйств, участвующих в измерении потребления энергии. Если наша гипотеза верна, то первыми (с наименьшими данными) будут праздничные дни.

	day	LCLid	energy_median	energy_mean	energy_max	energy_count	energy_std	energy_sum	energy_min
276	2013-10-04	4933	4933	4933	4933	4933	4933	4933	4933
345	2013-12-12	4947	4947	4947	4947	4947	4947	4947	4947
352	2013-12-19	4996	4996	4996	4996	4996	4996	4996	4996
338	2013-12-05	5010	5010	5010	5010	5010	5010	5010	5010
363	2013-12-30	5011	5011	5011	5011	5011	5011	5011	5011

Мы видим, что *дни с наименьшим количеством наблюдений не соответствуют праздничным дням*. Следовательно, *наше предположение*, что достаточным периодом наблюдения может являться количество дней в году за исключением праздничных - *неверно*.

Применим *другой метод оценки*. Построим гистограмму распределения по времени (по числу дней, в которые проводились измерения) количества домохозяйств, участвующих в измерениях.

Гистограмма распределения по времени количества домохозяйств участвующих в измерениях



Вывод: *большинство домохозяйств в 2013 году* имели период наблюдений (период измерений потребления электричества) **от 358 до 365 дней**. Интересно, что этот результат

численно совпал с нашим предыдущим предположением (о достаточности периода, состоящего из всех дней в году за исключением праздничных).

Посчитав, сколько всего домохозяйств **в 2013 году**, которые потребляли электроэнергию больше, чем **work_period (357)** дней, установили, что от исходного числа домохозяйств **(5528)** для дальнейшего анализа мы можем оставить только **4934** (занесем перечень этих домовладений в список ***list_hh_work_period***). Таким образом, количество "потерявшихся" домохозяйств теперь составляет **594.**

2.3. Финальное редактирование полученных данных

2.3.1. Приведем в порядок все рассматриваемые наборы данных: *удалим* "лишние" подмножества в соответствии с установленными нами окнами наблюдений.

Уменьшим выборку домохозяйств до заданного **% (<=100%)** . Для этого будем использовать разработанную нами *функцию*, которая будет возвращать выбранную случайным образом выборку длиной в заданный % в каждой **'Acorn_grouped'** от перечня домохозяйств заданного пересечением списков ***list_hh_NP_less*** и ***list_hh_work_period***.

Напомним, что *list_hh_NP_less* (содержит список домохозяйств из всех **'Acorn_grouped'**, кроме **'NP-Household'**, а *list_hh_work_period* (содержит список домохозяйств, которые передавали показания больше work_period (357**)** дней.

Для дальнейшего анализа возьмем 10% от всей выборки:

- `All Acorn_group summary: 4893`
- `Acorn_group: Adversity, summary: 1044, 10%: 92`
- `Acorn_group: Affluent, summary: 333, 10%: 30`
- `Acorn_group: Comfortable, summary: 1507, 10%: 137`
- `Acorn_group: Rising, summary: 1859, 10%: 161`
- `Acorn_group: Stretched, summary: 772, 10%: 70`

Всего: 490 домовладений

Из домохозяйств каждой группы взято рандомно десять процентов. Это будет *окончательный (рабочий) перечень домохозяйств*, которые оставляем для дальнейшего анализа электропотребления, назовем его ***list_hh_work***.

2.3.2. Сформируем рабочую выборку для набора данных **weather_daily_darksky** и **weather_hourly_darksky**

Из-за ошибки в наборе данных **weather_daily_darksky** *(по индексу 'time' отсутствует дата 2013-10-27, при этом дата 2013-03-31 «задвоена»)* предлагается в качестве **индекса 'time'** использовать **'sunriseTime'** обрезав время.

2.3.3. Сформируем рабочую выборки для наборов данных **daily_dataset** и **hhblock_dataset\block_XXX**.

Для формирования данной выборки по набору hhblock_dataset\block_XXX необходимо обработать в цикле часть файлов (из 112). Результатом выполнения будет созданный

датафрейм **data_hh_block_all**, который является объединением датафреймов, в которых значения столбца **'LCLid'** содержатся в списке **list_hh_work.**

Отметим, что общее число измерений смарт-счетчиков всех домохозяйств в рабочей выборке ***для hhblock_dataset\block_XXX*** совпало с аналогичным для ***daily_dataset***. Это факт подтверждает, что мы *корректно селектировали одни и те же домохозяйства и периоды наблюдений* для обоих наборов данных.

2.3.4. ***Переименуем*** **столбцы** датафрейма **data_hh_block_all** вида **hh_0, hh_1, hh_2** и т.д. (отражают получасовые периоды суток) в привычный вид с указанием времени суток **00:30:00, 01:00:00, 01:30:00** и т.д.

В результате получим таблицу следующего вида:

	LCLid	day	00:30:00	01:00:00	01:30:00	02:00:00	02:30:00	03:00:00	03:30:00	04:00:00	...	19:30:00	20:00:00	20:30:00	21:00:00	21:30:00
23620	MAC004387	2013-01-01	0.052	0.070	0.068	0.039	0.052	0.076	0.043	0.064	...	0.262	0.259	0.264	0.308	0.354
23621	MAC004387	2013-01-02	0.070	0.071	0.078	0.058	0.071	0.066	0.064	0.080	...	0.211	0.187	0.183	0.186	0.186
23622	MAC004387	2013-01-03	0.066	0.054	0.055	0.067	0.057	0.042	0.063	0.064	...	0.198	0.178	0.186	0.168	0.256
23623	MAC004387	2013-01-04	0.064	0.046	0.039	0.061	0.038	0.045	0.061	0.035	...	0.187	0.183	0.196	0.187	0.181
23624	MAC004387	2013-01-05	0.055	0.039	0.059	0.041	0.048	0.060	0.039	0.052	...	0.216	0.160	0.187	0.225	0.209

Общий вывод:

Наборы данных сформированы:

- **list_hh_work** - (рабочий) перечень домохозяйств, которые оставляем для дальнейшего анализа электропотребления;
- **data_daily_work_window** - выборка для набора данных **daily_dataset** (дневное энергопотребление каждого домохозяйства из **list_hh_work** *в интервале 2013.01.01 - 2013.12.31*);
- **data_weather_daily_work_window** - выборка для набора данных **weather_daily_darksky** *(в интервале 2013.01.01 - 2013.12.31)*;
- **data_weather_hourly_work_window** - выборка для набора данных **weather_hourly_darksky** *(в интервале 2013.01.01 - 2013.12.31)*;
- **data_hh_block_all** - выборка для набора данных **hhblock_dataset\block_XXX** (данные о получасовом энергопотребление каждого домохозяйства из **list_hh_work** в каждые сутки *в интервале 2013.01.01 - 2013.12.31*).

3. Анализ энергопотребления

3.1. Произведем анализ зависимости суммарного потребления энергии от среднесуточной температуры. Найдем средний прирост потребления энергии при изменении температуры

3.1.1. Построим график зависимости суммарного потребления энергии и среднесуточной температуры от времени

Сформируем набор данных **data_daily_energy,** включающий суммарное потребление энергии.

	day	energy_sum
0	2013-01-01	6069.982999
1	2013-01-02	5975.847000
2	2013-01-03	5703.976999
3	2013-01-04	5677.857002
4	2013-01-05	5869.862000

Сформируем набор данных **data_weather_d,** включающий информацию о среднедневной температуре.

	day	temperature_mean
427	2013-09-24	16.500
428	2013-07-26	20.725
429	2013-07-09	18.810
430	2013-08-08	17.820
431	2013-01-11	3.295

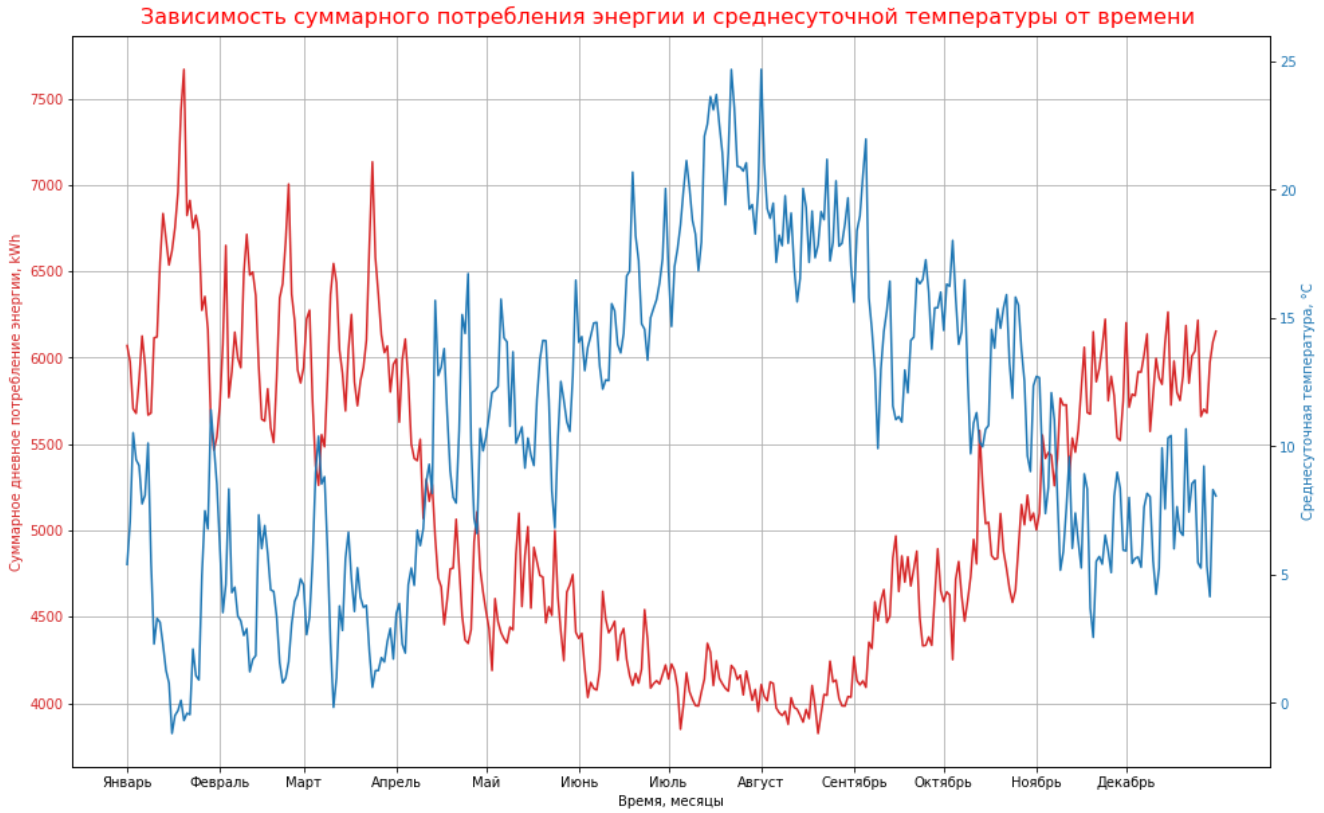
Объединим наборы данных **data_daily_energy** и **data_weather_d.**

	day	energy_sum	temperature_mean
0	2013-01-01	6069.982999	5.400
1	2013-01-02	5975.847000	7.065
2	2013-01-03	5703.976999	10.530
3	2013-01-04	5677.857002	9.480
4	2013-01-05	5869.862000	9.250

Добавим в набор данных **data_daily_energy** информацию о дате в **Unix формате**.

	day	energy_sum	temperature_mean	month
0	2013-01-01	6069.982999	5.400	1.356988e+09
1	2013-01-02	5975.847000	7.065	1.357074e+09
2	2013-01-03	5703.976999	10.530	1.357160e+09
3	2013-01-04	5677.857002	9.480	1.357247e+09
4	2013-01-05	5869.862000	9.250	1.357333e+09

Построим графики зависимости *суммарного потребления энергии и среднесуточной температуры от времени.*



На полученном графике хорошо видно, что кривые зависимости температур от времени года и потребленной электроэнергии от времени года практически зеркальны (по отношению друг другу) относительно оси, проходящей через ординату 10 градусов (по Цельсию) и находятся в явной связи. Исключение составляет отрезок, соответствующий периоду времени с июня по сентябрь.

3.1.2. Рассчитаем средний прирост потребления энергии при изменении среднедневной температуры на 1 градус Цельсия.

Выберем диапазоны данных для расчета показателей прироста/убыли среднего потребления энергии:

- **с Апреля по Июнь**, для расчета убыли среднего потребления энергии при повышении температуры,
- **с Сентября по Ноябрь**, для расчета прироста среднего потребления энергии при понижении температуры.

Процент **убыли** среднего потребления энергии **при увеличении** среднедневной температуры на 1 градус Цельсия **равен 2.28%.**

Процент **прироста** среднего потребления энергии **при уменьшении** среднедневной температуры на 1 градус Цельсия **равен 2.41%.**

Вывод: Совпадение (с точностью до погрешности) скорости изменения электропотребления при сезонном росте и понижении уличных температур позволяет нам сделать вывод не просто о связи эти двух величин, но и об их прямой зависимости. Далее мы еще вернемся к более подробному изучению этой зависимости и построим математическую модель, связывающую температурные изменения и потребление электроэнергии.

3.2. Сделаем сравнение потребления энергии в выходные/праздничные и рабочие дни

Сформируем набор данных **data_daily_Acorn_grouped** из двух таблиц **data_daily_work_window** и **data_informations_households** по ключу *LCLid*

	LCLid	day	energy_median	energy_mean	energy_max	energy_count	energy_std	energy_sum	energy_min	stdorToU	Acorn	Acorn_grouped
0	MAC000235	2013-01-01	0.0315	0.030979	0.064	48	0.011650	1.487	0.012	Std	ACORN-E	Rising
1	MAC000235	2013-01-02	0.0320	0.034646	0.206	48	0.027834	1.663	0.011	Std	ACORN-E	Rising
2	MAC000235	2013-01-03	0.0290	0.027625	0.050	48	0.008746	1.326	0.013	Std	ACORN-E	Rising
3	MAC000235	2013-01-04	0.0295	0.027729	0.047	48	0.008434	1.331	0.013	Std	ACORN-E	Rising
4	MAC000235	2013-01-05	0.0305	0.027750	0.045	48	0.008760	1.332	0.012	Std	ACORN-E	Rising

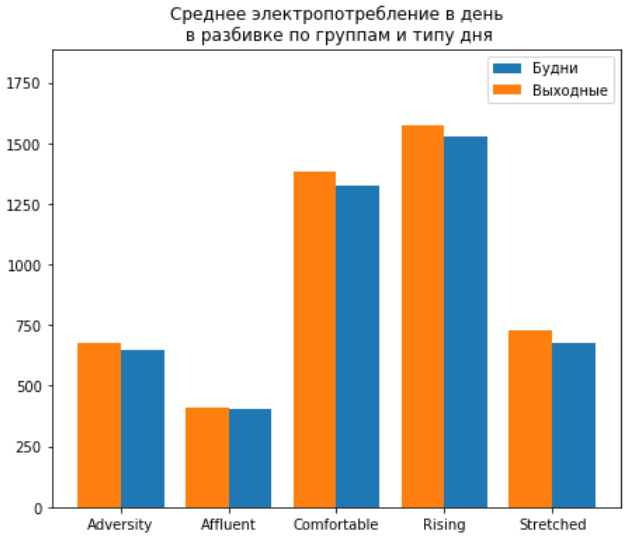
Сформируем набор данных **data_daily_grouped** включающий информацию о суммарном потреблении энергии, количестве домохозяйств, вычислим и добавим в него информацию о будних, выходных днях и о потреблении энергии по категориям будний/выходной день.

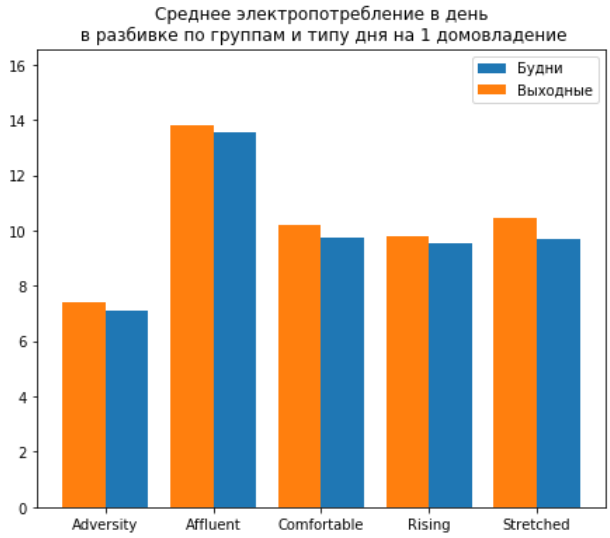
	day	Acorn_grouped	energy_sum	LCLid	holiday	workday	workday_all	holiday_all	workday_house	holiday_house
0	2013-01-01	Adversity	745.645000	92	1	0	0.0	745.645000	0.0	8.104837
1	2013-01-01	Affluent	491.727000	30	1	0	0.0	491.727000	0.0	16.390900
2	2013-01-01	Comfortable	1597.005998	137	1	0	0.0	1597.005998	0.0	11.656978
3	2013-01-01	Rising	1715.421999	161	1	0	0.0	1715.421999	0.0	10.654795
4	2013-01-01	Stretched	875.399001	70	1	0	0.0	875.399001	0.0	12.505700

Сформируем набор данных **for_plot_mean_workweek**, включающий информацию о потребленной энергии в будние/выходные дни по группам потребителей. Добавим в него столбцы для построения гистограмм: суммарная средняя потребленная энергия в будние/выходные дни, средняя потребленная энергия на 1 домохозяйство в будние/выходные дни.

	Acorn_grouped	workday_all	holiday_all	workday_house	holiday_house
0	Adversity	649.850285	677.277768	7.105440	7.407543
1	Affluent	404.306308	412.319250	13.556719	13.806792
2	Comfortable	1328.269743	1385.791554	9.747773	10.178169
3	Rising	1525.718419	1574.533152	9.528210	9.818326
4	Stretched	675.002763	728.346875	9.687149	10.444726

Построим гистограммы суммарного потребления по каждой из групп потребителей.





На представленных графиках видно, что*:*

- общее энергопотребление (и в выходные и в рабочие дни), главным образом, зависит от числа домовладений в каждой из групп, в представленной выборке
- энергопотребление в выходные дни превышает энергопотребление в рабочие дни
- в ранжировании энергопотребления в расчете на одно домовладение с большим отрывом вырывается вперед самая обеспеченная часть. Далее, практически "вровень", идут средний класс, и, также, с отрывом, но в вниз, наименее обеспеченные.

Для каждой группы потребителей найдем прирост среднесуточного потребления энергии в рабочие дни к потреблению в выходные дни.

	Acorn_grouped	workday_all	holiday_all	workday_house	holiday_house	att
1	Affluent	404.306308	412.319250	13.556719	13.806792	1.81
3	Rising	1525.718419	1574.533152	9.528210	9.818326	2.95
0	Adversity	649.850285	677.277768	7.105440	7.407543	4.08
2	Comfortable	1328.269743	1385.791554	9.747773	10.178169	4.23
4	Stretched	675.002763	728.346875	9.687149	10.444726	7.25

Построим диаграмму:

Общий вывод по данной части:

- потребление электроэнергии в выходные дни возрастает, по сравнению с будними днями
- наименьший прирост потребления электроэнергии в выходные дни наблюдается в обеспеченной группе. При этом явно не прослеживается зависимость величины прироста от финансового благополучия

3.3. Для каждой из пяти групп потребителей проведем анализ суточного энергопотребления в течение 2013 года

- суммарно на группу и в пересчете на одно домохозяйство с учетом светового дня, и
- отдельно по тарифным группам: **T1** - пиковая нагрузка, **T2** - средняя нагрузка, **T3** - низкая нагрузка.

Сделаем выводы о периодах суток с наибольшим энергопотреблением, и о влиянии на энергопотреблением выбранного тарифа в каждой из пяти групп.

Определим периоды T1, T2, T3, в соответствии с общепринятой практикой:

- T1 - с 07:00 до 10:00 и с 17:00 до 21:00;
- T2 - с 10:00 до 17:00 и с 21:00 до 23:00;
- T3 - с 23:00 до 24:00 и с 00:00 до 07:00.

Добавим в набор **data_hh_block_all** данные о времени восхода и заката солнца, присвоив значение **data_hh_block_all_sun:**

	LCLid	day	00:30:00	01:00:00	01:30:00	02:00:00	02:30:00	03:00:00	03:30:00	04:00:00	...	sunriseTime	sunsetTime	day_power	night_power
0	MAC003388	2013-01-01	1.600	1.580	1.518	1.424	1.952	1.531	1.110	0.770	...	2013-01-01 08:07:21	2013-01-01 16:03:05	17.884	18.750
1	MAC003388	2013-01-02	0.258	0.219	0.223	0.146	0.149	0.205	0.200	0.261	...	2013-01-02 08:07:12	2013-01-02 16:04:09	2.473	6.589
2	MAC003388	2013-01-03	0.202	0.308	0.240	0.140	0.096	0.134	0.161	0.095	...	2013-01-03 08:07:00	2013-01-03 16:05:15	7.467	17.269
3	MAC003388	2013-01-04	0.663	0.704	0.468	0.130	0.152	0.172	0.212	0.162	...	2013-01-04 08:06:45	2013-01-04 16:06:24	6.836	18.910
4	MAC003388	2013-01-05	0.321	0.271	0.140	0.159	0.172	0.164	0.240	0.273	...	2013-01-05 08:06:26	2013-01-05 16:07:35	14.894	22.934

Используем разработанную нами функцию, которая вычисляет *за одни сутки для одного домохозяйства* дневное (от восхода до заката), ночное и суммарное *потребление электроэнергии*, а также *потребление энергии в тарифные интервалы* **T1, T2, T3**.

Вычислим для всех домохозяйств *дневное, ночное и суммарное* потребление электроэнергии.

Добавим к полученному набору данных **data_hh_block_all_sun** информацию о принадлежности домохозяйств к одной из пяти групп **Acorn_grouped**.

В результате получим таблицу вида:

03:00:00	03:30:00	04:00:00	...	sunriseTime	sunsetTime	day_power	night_power	summary_power	T1_power	T2_power	T3_power	Acorn_grouped	stdorToU
1.531	1.110	0.770	...	2013-01-01 08:07:21	2013-01-01 16:03:05	17.884	18.750	36.634	7.184	15.695	13.755	Affluent	Std
0.205	0.200	0.261	...	2013-01-02 08:07:12	2013-01-02 16:04:09	2.473	6.589	9.062	2.801	3.307	2.954	Affluent	Std
0.134	0.161	0.095	...	2013-01-03 08:07:00	2013-01-03 16:05:15	7.467	17.269	24.736	8.309	12.523	3.904	Affluent	Std
0.172	0.212	0.162	...	2013-01-04 08:06:45	2013-01-04 16:06:24	6.836	18.910	25.746	11.069	8.920	5.757	Affluent	Std
0.164	0.240	0.273	...	2013-01-05 08:06:26	2013-01-05 16:07:35	14.894	22.934	37.828	14.840	18.608	4.380	Affluent	Std

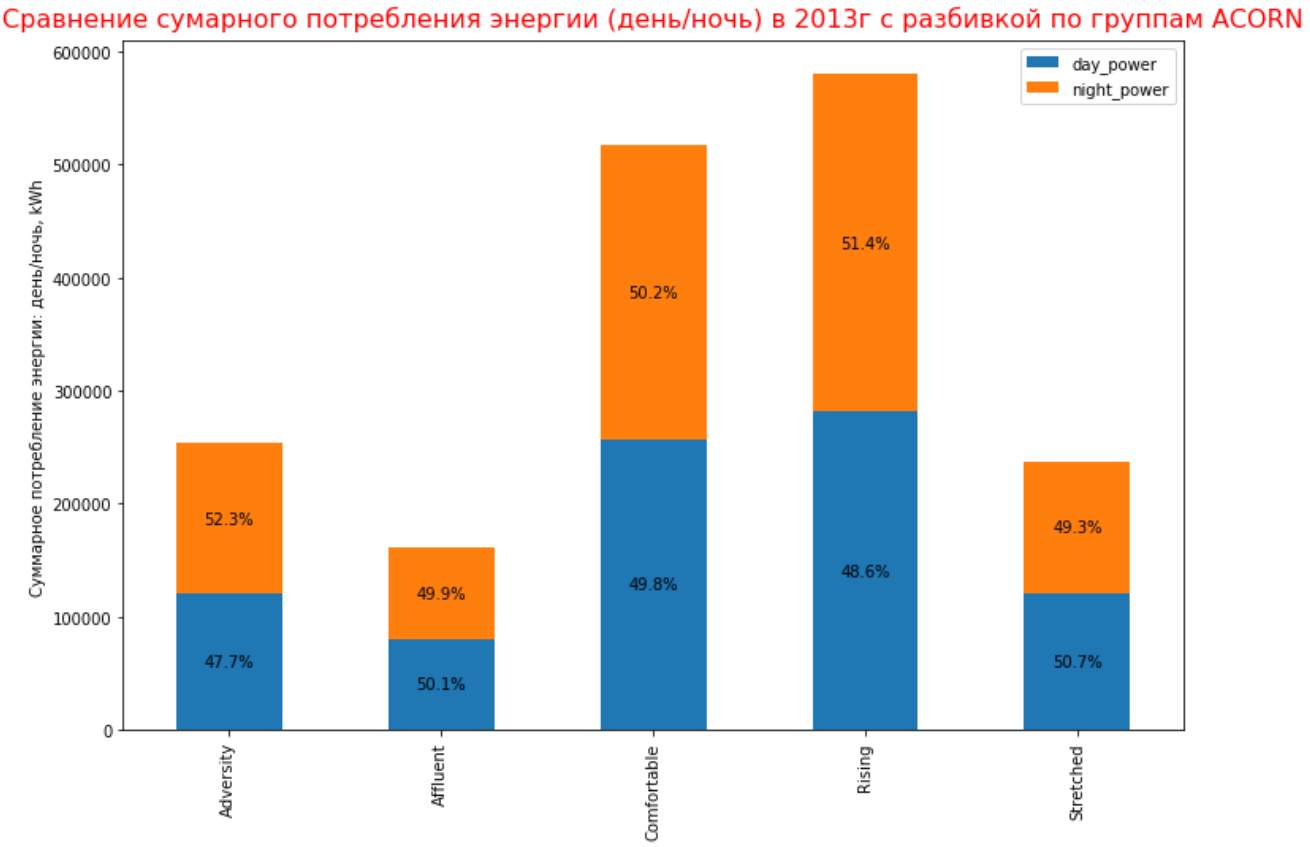
3.3.1. Сравнение ночного, дневного и суммарного энергопотребления

Для каждой из пяти групп потребителей *сравним ночное/дневное/суммарное потребление* суммарно на группу и в пересчете на одно домохозяйство. Построим диаграммы.

Сгруппируем все домохозяйства по принадлежности к **Acorn_grouped**, и для каждой группы *посчитаем суммарные значения дневного, ночного и суточного потребления электроэнергии.*

Acorn_grouped	LCLid	day_power	night_power	summary_power
Adversity	92	120817.359998	132638.441999	253455.801997
Affluent	30	80934.210997	80694.020998	161628.231995
Comfortable	137	257131.223992	259454.882990	516586.106982
Rising	161	281601.103996	298343.962981	579945.066978
Stretched	70	120411.353998	117115.605007	237526.959005

Построим **столбиковую диаграмму** *суммарного потребления энергии (день/ночь) в 2013 году* с разбивкой по группам **ACORN.**

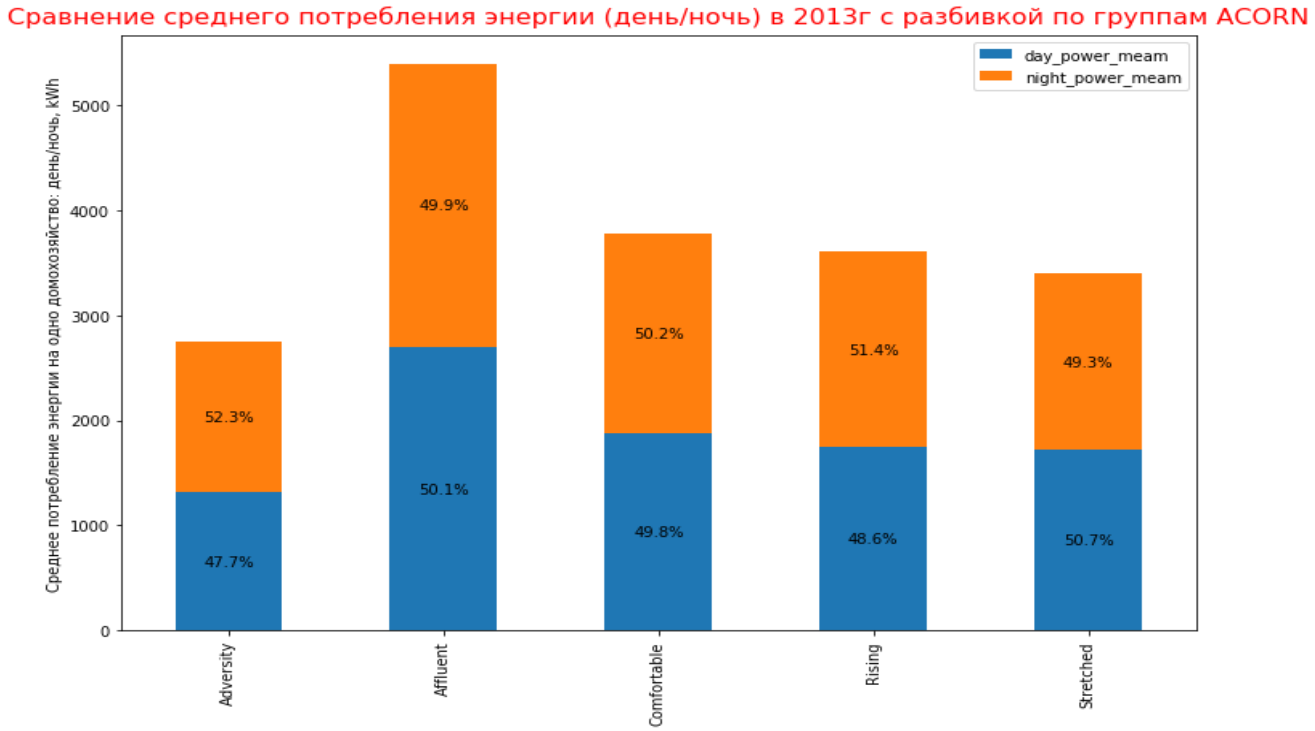


Построим **круговую диаграмму**, показывающую *распределение долей электроэнергии каждой из групп в суммарном дневном и в суммарном ночном потреблении.*

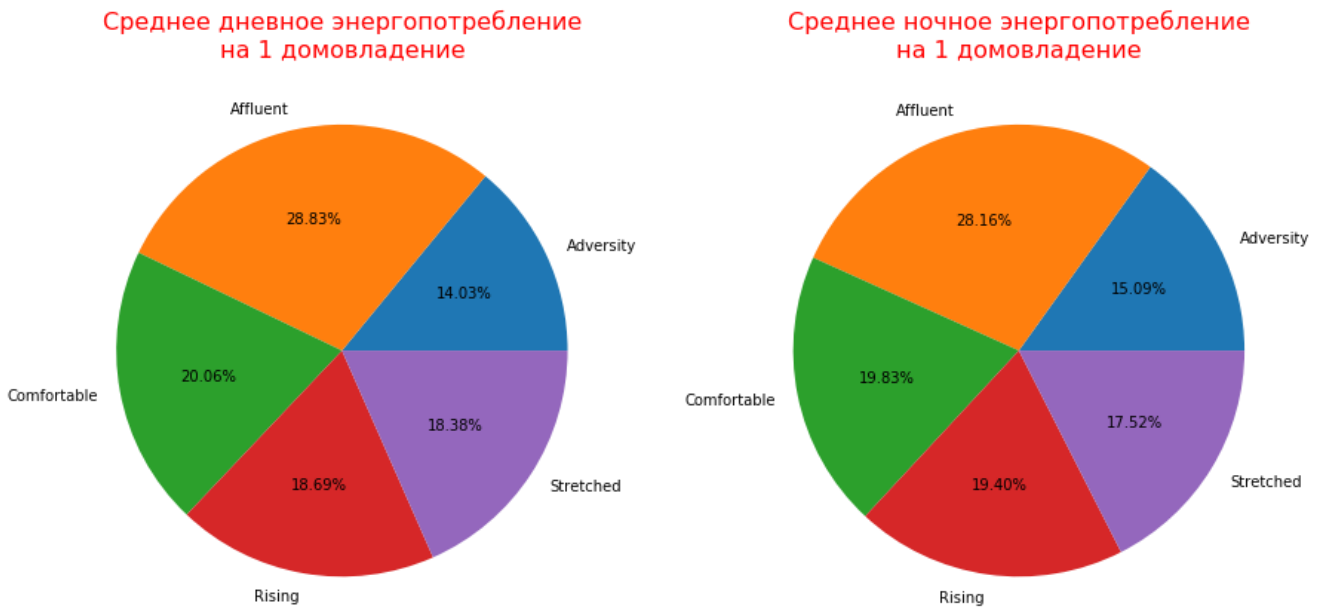
Рассчитаем **среднее дневное потребление** (на одно домохозяйство) для каждой из групп.

Acorn_grouped	LCLid	day_power	night_power	summary_power	day_power_meam	night_power_meam
Adversity	92	120817.359998	132638.441999	253455.801997	1313.232174	1441.722196
Affluent	30	80934.210997	80694.020998	161628.231995	2697.807033	2689.800700
Comfortable	137	257131.223992	259454.882990	516586.106982	1876.870248	1893.831263
Rising	161	281601.103996	298343.962981	579945.066978	1749.075180	1853.068093
Stretched	70	120411.353998	117115.605007	237526.959005	1720.162200	1673.080072

Построим **столбиковую диаграмму** *среднего потребления энергии (день/ночь) в 2013 году* с разбивкой по группам **ACORN.**



Построим **круговую диаграмму**, показывающую *распределение долей энергопотребления в среднем на одно домовладение в каждой из групп в дневное и в ночное время.*



Представленные диаграммы наглядно показывают, что:

- суммарное энергопотребление, в каждой из групп домовладений, определяется ее численностью. Этот факт мы уже отмечали выше
- доли дневного и ночного потребления электроэнергии (в светлое и темное время суток) в среднегодовом измерении практически равны друг другу во всех группах домовладений
- выявление факта, что доли дневного/ночного энергопотребления равны примерно 50% во всех группах домовладений, исключает возможность его дополнительного ранжирования в зависимости от принадлежности домовладения к той или иной группе. Можем лишь констатировать, уже ранее установленный результат, что в ранжировании энергопотребления в расчете на одно домовладение с большим отрывом вырывается вперед самая обеспеченная часть. Далее, с минимальным градиентом на понижение идут средний клас и менее обеспеченные.

3.3.2. Сравнение энергопотребления по периодам T1, T2 и T3

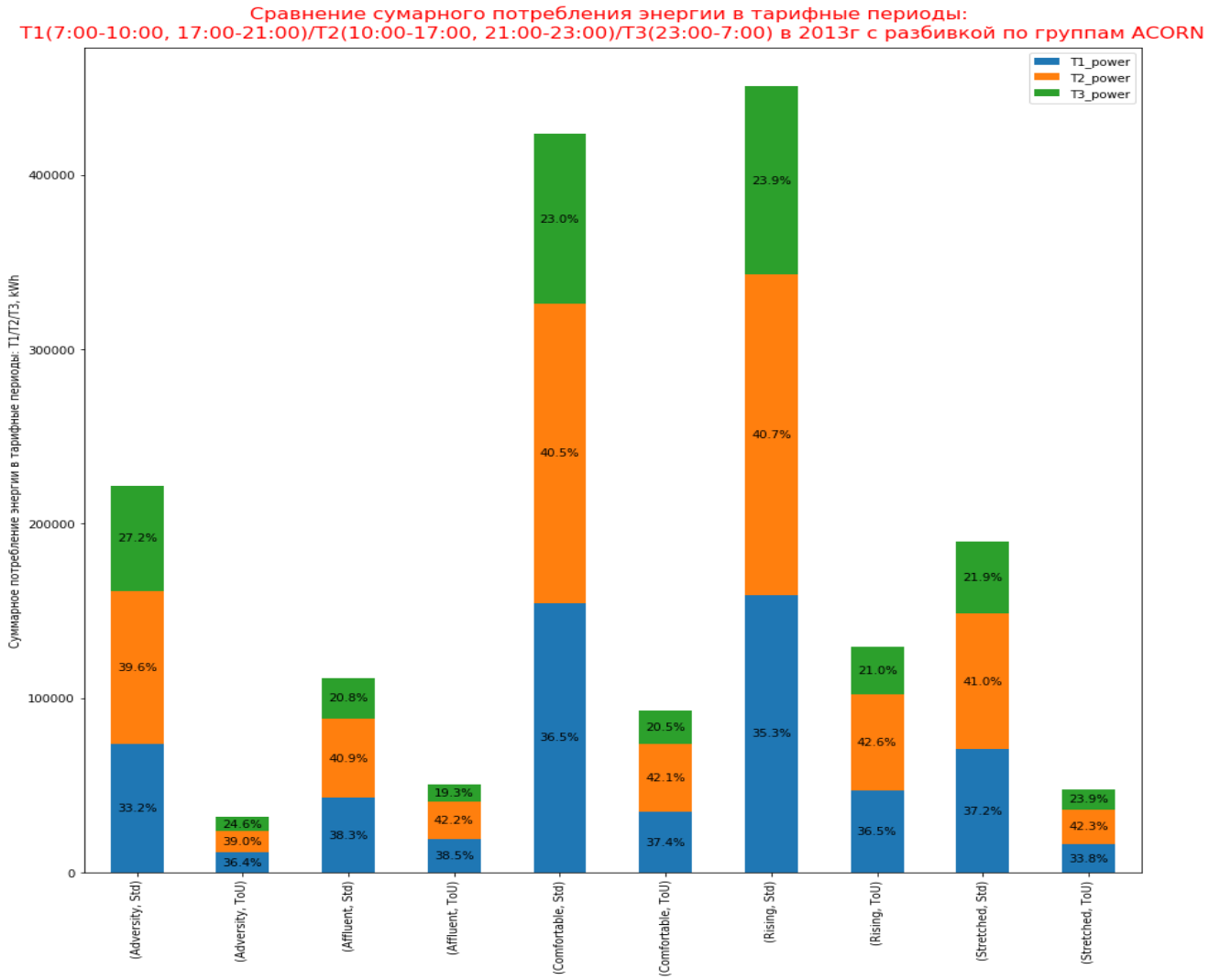
Для каждой из *пяти групп потребителей* сравним энергопотребление (суммарно на группу и в пересчете на одно домохозяйство) с учетом выбранного тарифа в периоды: **T1** - *пиковая нагрузка*, **T2** - *средняя нагрузка*, **T3** - *низкая нагрузка*.

Сделаем выводы о периодах суток с наибольшим энергопотреблением, и о влиянии на энергопотребление выбранного тарифа в каждой из пяти групп.

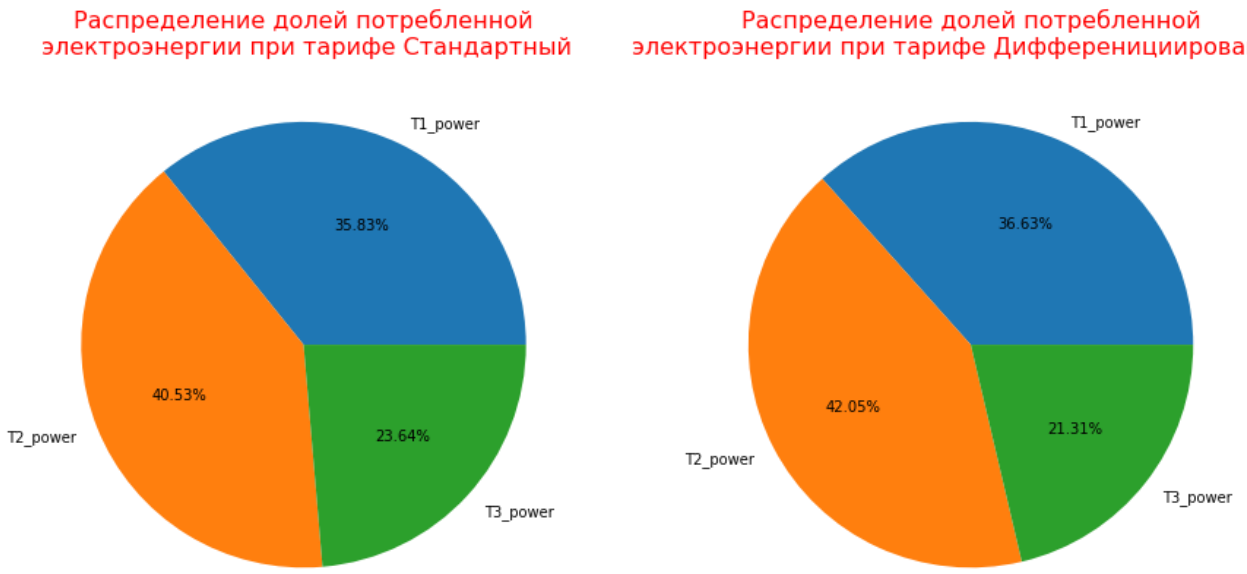
Сгруппируем все домохозяйства по принадлежности к **Acorn_grouped** и выбранному тарифу **stdorToU**, и для каждой тарифной подгруппы посчитаем суммарные значения потребления электроэнергии в периоды **T1, T2, T3.**

Acorn_grouped	stdorToU	LCLid	summary_power	T1_power	T2_power	T3_power
Adversity	Std	77	221807.685997	73652.771999	87892.086994	60262.827004
	ToU	15	31648.116000	11516.223000	12352.867000	7779.026000
Affluent	Std	19	111446.832994	42700.840993	45560.250997	23185.741004
	ToU	11	50181.399000	19320.966003	21194.160998	9666.272000
Comfortable	Std	107	423599.855978	154497.560994	171534.749990	97567.544993
	ToU	30	92986.251005	34747.508002	39189.645003	19049.097999
Rising	Std	121	450585.347980	159195.811990	183523.538999	107865.996992
	ToU	40	129359.718997	47180.379998	55074.712997	27104.626002
Stretched	Std	58	189971.756006	70580.961005	77873.576999	41517.218002
	ToU	12	47555.202999	16090.571998	20094.087001	11370.544001

Построим **столбиковую диаграмму** *суммарного потребления энергии в тарифные периоды:* **T1**(7:00-10:00, 17:00-21:00)/**T2**(10:00-17:00, 21:00-23:00)/**T3**(23:00-7:00) **в 2013 году** с разбивкой по группам **ACORN**



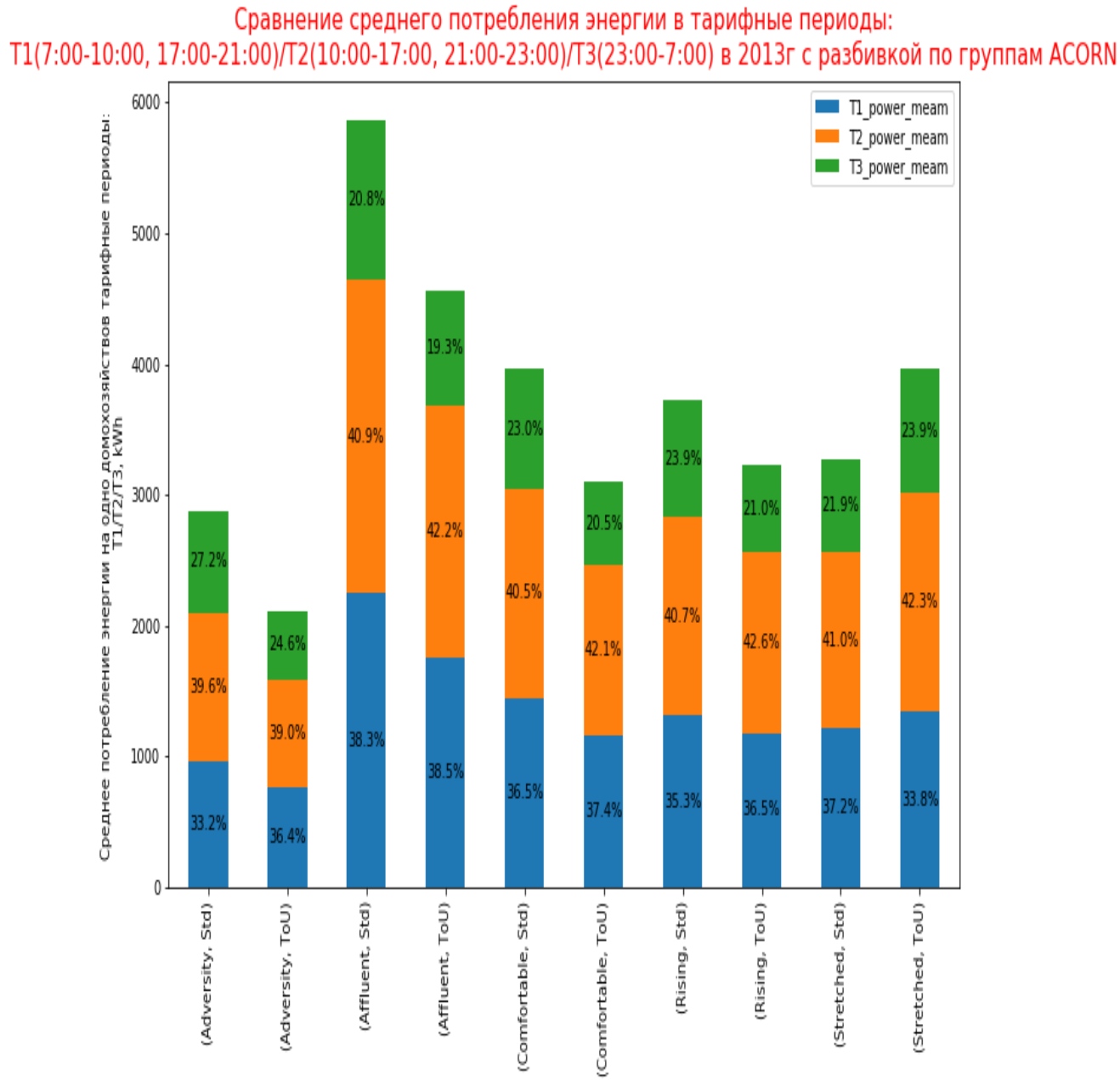
Построим **круговую диаграмму**, показывающую *распределение долей потребленной электроэнергии по периодам с пиковой, средней, и минимальной нагрузкой* отдельно для всех *домовладений* со стандартным тарифом и дифференцированным тарифом.



Рассчитаем среднее потребление энергии на одно домохозяйство в тарифные периоды **T1/T2/T3** для каждой из групп.

Acorn_grouped	stdorToU	LCLid	summary_power	T1_power	T2_power	T3_power	T1_power_meam	T2_power_meam	T3_power_meam
Adversity	**Std**	77	221807.685997	73652.771999	87892.086994	60262.827004	956.529506	1141.455675	782.634117
	ToU	15	31648.116000	11516.223000	12352.867000	7779.026000	767.748200	823.524467	518.601733
Affluent	**Std**	19	111446.832994	42700.840993	45560.250997	23185.741004	2247.412684	2397.907947	1220.302158
	ToU	11	50181.399000	19320.966003	21194.160998	9666.272000	1756.451455	1926.741909	878.752000
Comfortable	**Std**	107	423599.855978	154497.560994	171534.749990	97567.544993	1443.902439	1603.128505	911.846215

Построим **столбиковую диаграмму** *среднего потребления энергии* на одно домохозяйство в тарифные периоды **T1/T2/T3 в 2013 году** с разбивкой по группам **ACORN.**



Построим **круговую диаграмму**, показывающую *распределение долей потребленной электроэнергии* в среднем на одно домовладение *со стандартным тарифом и дифференцированным тарифом* в каждой из групп - отдельно *в периоды с пиковой, средней, и минимальной нагрузкой.*

Выводы:

- сравнение суммарного потребления электроэнергии всех домовладений с разбивкой по тарифным периодам показывает, что:
  - максимальное потребление наблюдается в период со средней нагрузкой (Т2), а минимальное потребление - в период с низкой нагрузкой (Т3)
  - при этом выбор тарифа (стандартный или дифференцируемый) не оказывает значительного влияния на потребление ни в одном из периодов Т1, Т2 или Т3. Т.е., если не принимать во внимание принадлежность домовладений к группам (Acorn), то можно сказать, что выбор дифференцированного тарифа в среднем по всем группам не приводит к экономии электроэнергии. В противном случае мы бы наблюдали увеличение доли потребления в период Т3 и уменьшения долей в Т1 и Т2, у домовладений с дифференцированным тарифом
- сравнение долей среднего потребления электроэнергии одного домовладения для каждой из групп с разбивкой по тарифным периодам показывает, что:

- домовладения с дифференцированным тарифом за сутки, как правило, потребляют меньше электроэнергии, чем со стандартным тарифом. Однако эта разница достигается не за счет перераспределения потребления в пользу периода с низкой нагрузкой (и минимальным тарифом) - Т3
- как и случае с общим энергопотреблением, ни для одной из групп домовладений мы не можем сказать, что выбор дифференцированного тарифа приводит к экономии электроэнергии за счет увеличения доли потребления в период Т3 и уменьшения долей в Т1 и Т2.

3.4. Для каждой из пяти групп потребителей построим графики зависимости суточного потребления энергии от длины светового дня

Совместим их на одной координатной плоскости с графиками зависимости суточного потребления энергии от среднесуточной температуры комфорта. Найдем средний прирост потребления энергии при увеличении температуры и сокращении длины светового дня в каждой из пяти групп в пересчете на один час.

3.4.1. Построим графики зависимости суточного потребления энергии от длины светового дня, и от среднесуточной температуры.

Сформируем набор данных **data_daily_graph,** включающий информацию о группе потребителей и среднем дневном потреблении энергии.

	day	Acorn_grouped	energy_sum
0	2013-01-01	Adversity	9.040304
1	2013-01-01	Affluent	19.295767
2	2013-01-01	Comfortable	12.356599
3	2013-01-01	Rising	10.390662
4	2013-01-01	Stretched	11.677443

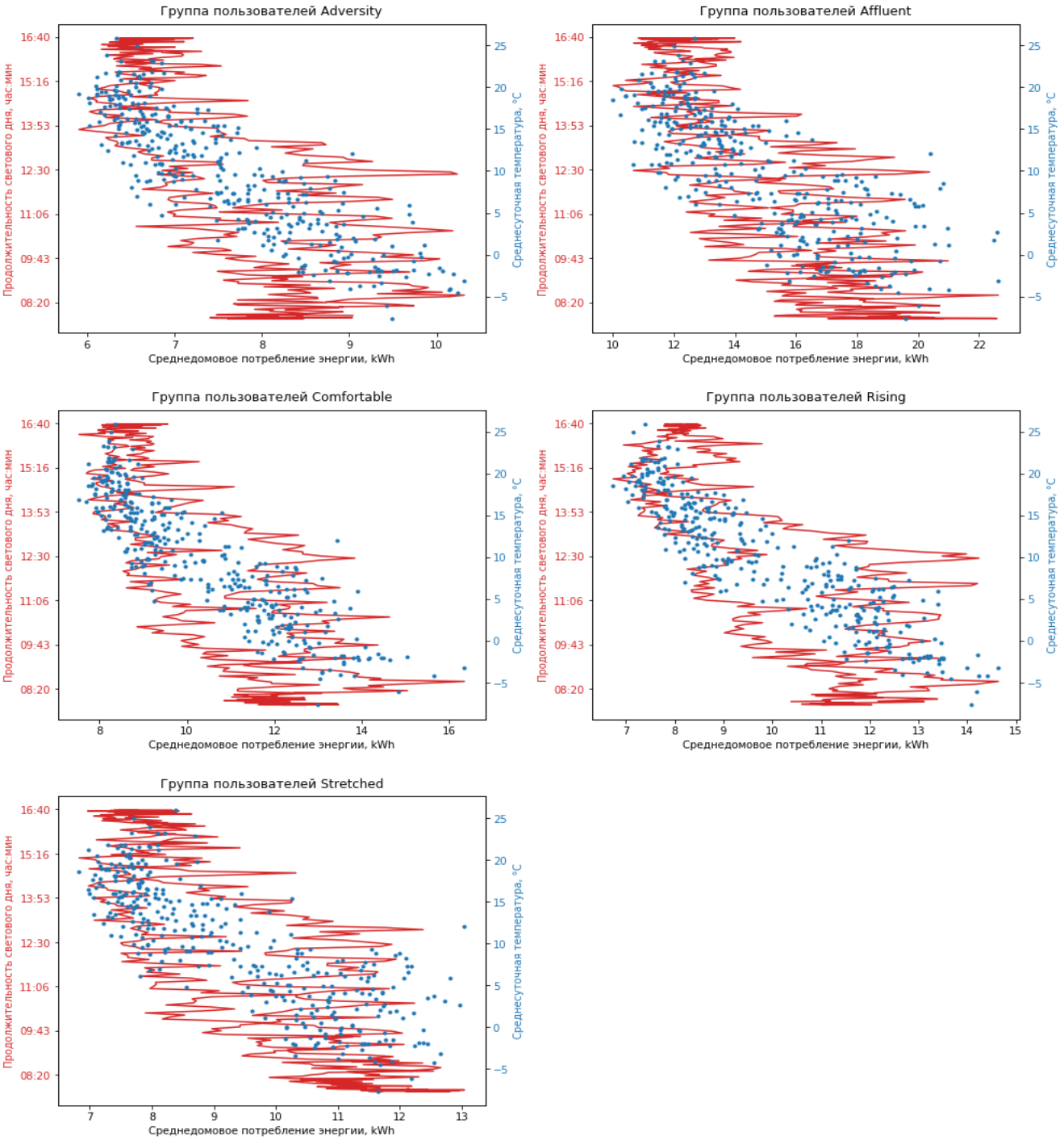
Создадим набор данных **data_weather_daily_graph** с информацией о длине светового дня и среднедневной температуре комфорта.

	day	time_sun_unix	apparentTemperature
427	2013-09-24	43513.0	17.570
428	2013-07-26	56574.0	20.505
429	2013-07-09	58949.0	19.575
430	2013-08-08	54122.0	19.765
431	2013-01-11	29524.0	2.205

Объединим наборы данных **data_daily_graph** и **data_weather_daily_graph**.

	day	Acorn_grouped	energy_sum	time_sun_unix	apparentTemperature
0	2013-01-01	Adversity	9.040304	28544.0	2.01
1	2013-01-01	Affluent	19.295767	28544.0	2.01
2	2013-01-01	Comfortable	12.356599	28544.0	2.01
3	2013-01-01	Rising	10.390662	28544.0	2.01
4	2013-01-01	Stretched	11.677443	28544.0	2.01

Построим графики зависимости среднесуточного потребления энергии от длины светового дня по группам **ACORN**.



Из представленных графиков видно, что:

- влияние, которое оказывает на потребление электроэнергии, значение среднесуточных температур пропорционально влиянию продолжительности светового дня
- эта пропорция сохраняется для домовладений из всех пяти групп

3.4.2. Рассчитаем средний прирост потребления энергии при изменении среднедневной температуры комфорта и длительности светового дня.

Рассчитаем коэффициенты:

- **avg_temp**, равен количеству интервалов 5 градусов в диапазоне среднедневной температуры комфорта.
- **avg_time**, равен количеству 1 часовых интервалов в диапазоне светового дня.

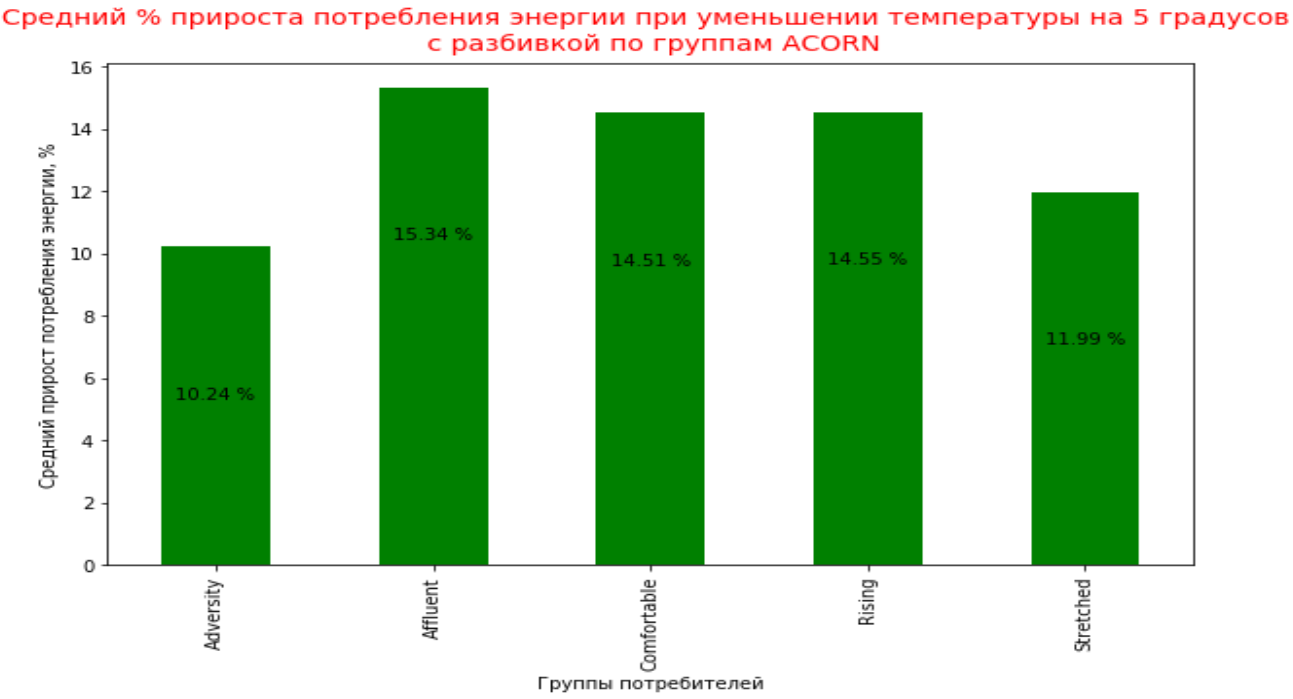
По каждой из групп потребления рассчитаем *средние темпы прироста* потребления энергии (в %):

1. при уменьшении температуры на 5 градусов.
2. при сокращении длины светового дня на 1 час.

Acorn_grouped	t_mean_time	t_mean_temp
Adversity	7.38	10.24
Affluent	10.99	15.34
Comfortable	10.41	14.51
Rising	10.44	14.55
Stretched	8.63	11.99

*Построим **диаграмму прироста*** потребления энергии *при уменьшении **светового дня** на 1 час* с разбивкой по группам **ACORN.**

Построим ***диаграмму прироста*** потребления *энергии* при уменьшении **среднедневной *температуры*** *комфорта на 5 градусов* с разбивкой по группам **ACORN**.



Выводы:

- изменение значений температуры комфорта на 5 градусов (Цельсия) сопоставимо по влиянию на энергопотребление изменения длины светового дня на 1 час (примерно как 1:1.3)
- сильнее всего влияние обоих факторов сказывается на энергопотреблении самой обеспеченной группы и далее с незначительным градиентом оно уменьшается с понижением уровня благосостояния групп

3.5. Исследуем зависимость суммарного потребления энергии в течение года от времени суток; от дня недели и месяца года.

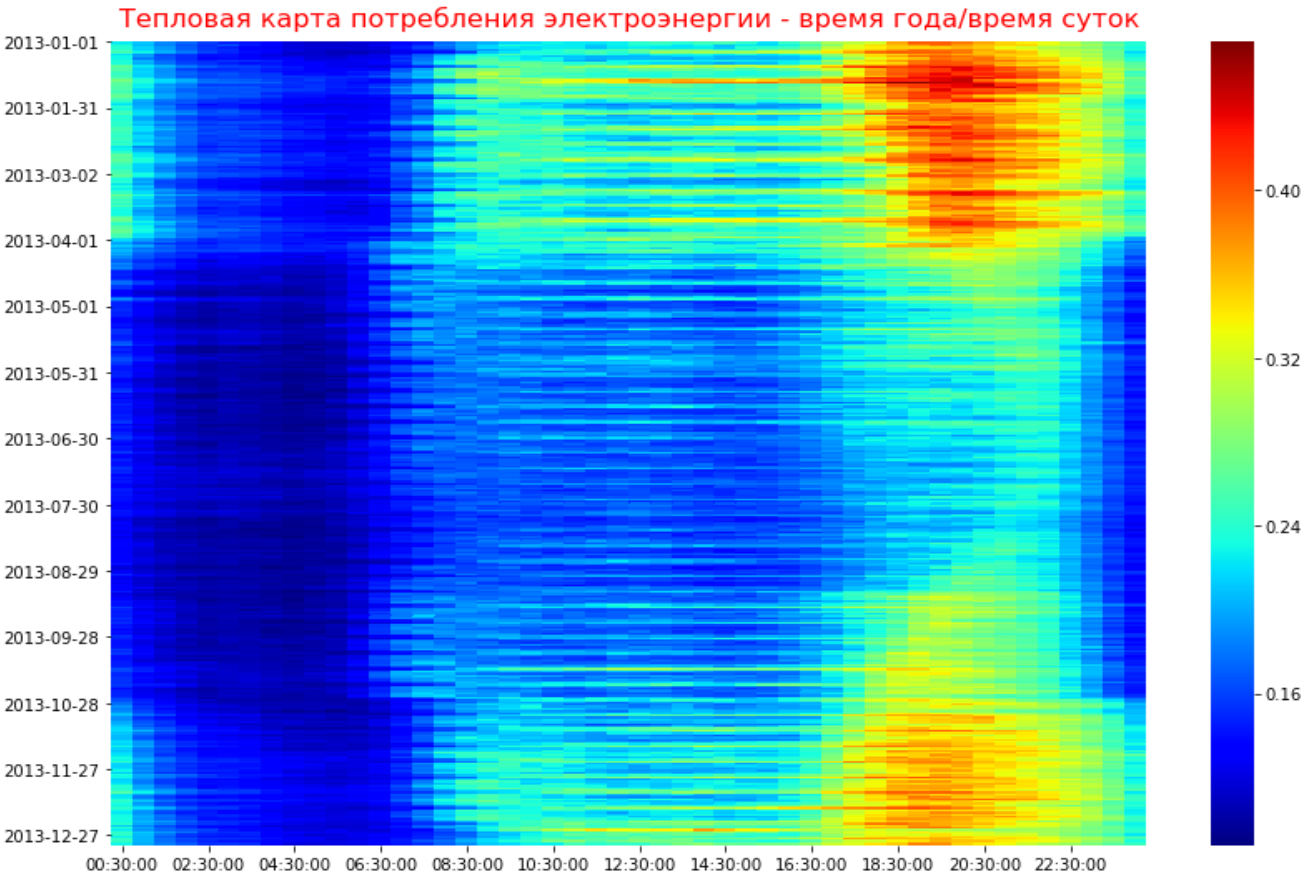
Построим качественную оценку динамики энергопотребления с помощью тепловых карт: спектр потребления энергии от темно-синего (мин.) до темно-красного (макс)

3.5.1. Тепловая карта потребления электроэнергии - время года/время суток: по осям дни года (ось Y) и часы суток (ось X).

При формировании набора *данных для тепловой карты* будем использовать выборку **data_hh_block_all**. Сгруппируем данный набор данных по ключу **'day'** и найдем *средние значения от потребления электроэнергии всеми домовладениями в течение каждых суток года*. Запишем этот набор данных в переменной **data_heatmap_all**, которую и будем использовать при построении графика тепловой карты.

day	00:30:00	01:00:00	01:30:00	02:00:00	02:30:00	03:00:00	03:30:00	04:00:00	04:30:00	05:00:00	...	19:30:00	20:00:00	20:30:00
2013-01-01	0.257307	0.229798	0.210926	0.200992	0.189155	0.172080	0.162104	0.145190	0.126008	0.117170	...	0.368272	0.377724	0.354065
2013-01-02	0.224328	0.210334	0.186746	0.173949	0.158238	0.147670	0.140531	0.130467	0.118750	0.113150	...	0.397941	0.387070	0.369061
2013-01-03	0.234761	0.205545	0.183651	0.165761	0.153249	0.168308	0.147231	0.132833	0.119059	0.115478	...	0.392110	0.371286	0.356651
2013-01-04	0.239271	0.213811	0.193140	0.171431	0.155735	0.146207	0.134864	0.128111	0.118678	0.116446	...	0.352700	0.364624	0.355873
2013-01-05	0.243592	0.214537	0.196605	0.172250	0.155154	0.152855	0.135289	0.132568	0.123617	0.115539	...	0.385098	0.386236	0.370592

Построим **тепловую карту** энергопотребления:



Данная тепловая карта иллюстрирует хотя и вполне ожидаемый, но весьма нагдяный результат:

- самое большое потребление электроэнергии в течении года происходит в вечерние часы с 17:00 24:00
- на втором месте по потреблению дневные часы с 9:00 до 17:00
- пиковые значения приходятся на часы с 18:30 до 22:00 в период с конца октября по начало апреля

3.5.2. Тепловая карта потребления электроэнергии - дни недели/недели года: по осям неделя года (ось Y) и дни недели (ось X).

При формировании набора данных для тепловой карты будем использовать выборку **data_daily_work_window.** Создадим новый датафрейм, добавив в него информацию о дне недели и порядковом номере недели в году:

	LCLid	day	energy_sum	weekday	weekofyear
6123	MAC004356	2013-01-01	11.947	1	1
6124	MAC004356	2013-01-02	8.704	2	1
6125	MAC004356	2013-01-03	10.385	3	1
6126	MAC004356	2013-01-04	9.631	4	1
6127	MAC004356	2013-01-05	12.997	5	1

Затем, *для каждой строки* определим **день недели** и запишем *потребленную в этот день электроэнергию* в соответствующий столбец:

	LCLid	day	energy_sum	weekday	weekofyear	Mon	Tue	Wed	Thu	Fri	Sat	Sun
6123	MAC004356	2013-01-01	11.947	1	1	NaN	11.947	NaN	NaN	NaN	NaN	NaN
6124	MAC004356	2013-01-02	8.704	2	1	NaN	NaN	8.704	NaN	NaN	NaN	NaN
6125	MAC004356	2013-01-03	10.385	3	1	NaN	NaN	NaN	10.385	NaN	NaN	NaN
6126	MAC004356	2013-01-04	9.631	4	1	NaN	NaN	NaN	NaN	9.631	NaN	NaN
6127	MAC004356	2013-01-05	12.997	5	1	NaN	NaN	NaN	NaN	NaN	12.997	NaN
6128	MAC004356	2013-01-06	12.082	6	1	NaN	NaN	NaN	NaN	NaN	NaN	12.082

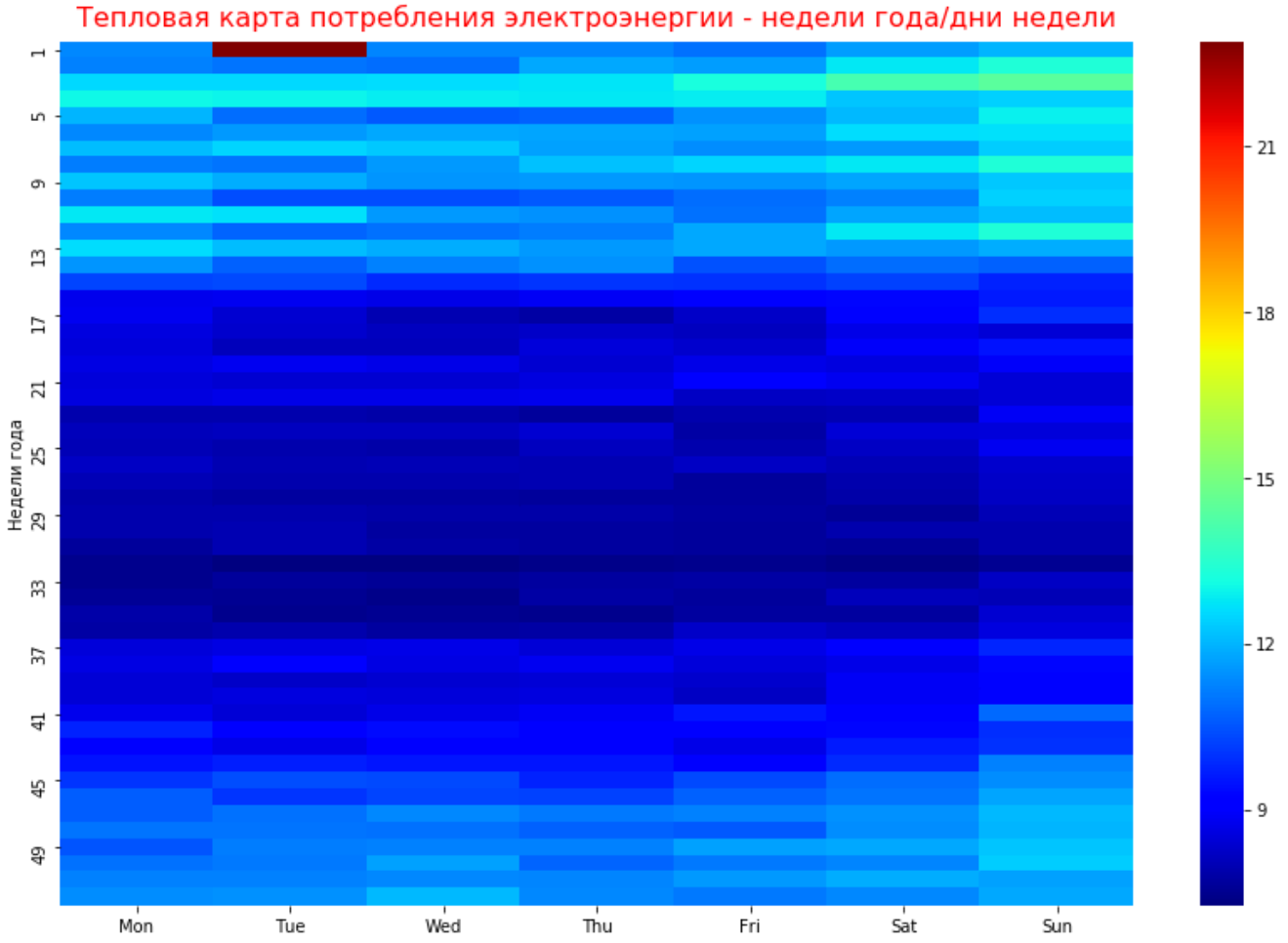
Удалим не нужные нам столбцы **'day', 'energy_sum', 'weekday'**. *Сгруппируем* по составному **ключу 'weekofyear', 'LCLid'** и *просуммируем* значения столбцов**: 'Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun'**. Получим в каждой строке значения потребления электроэнергии одного домовладения в течение одной недели.

Затем *сгруппируем* полученный набор по **ключу 'weekofyear'** и *найдем средние значения* от потребления электроэнергии всеми домовладениями в течение каждой недели года.

Запишем этот набор данных в переменной **data_heatmap_week**, которую и будем использовать при построении графика тепловой карты.

weekofyear	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	11.337871	22.892033	11.266610	11.251941	10.984394	11.636204	11.982204
2	11.224392	11.025022	10.906145	11.792722	11.625680	12.770680	13.315353
3	12.521092	12.530055	12.629029	12.753598	13.233673	14.089922	14.457624
4	13.023806	12.987292	12.840745	12.765208	12.865192	12.214465	12.413963
5	12.010324	10.883094	10.571341	10.737304	11.451684	12.072610	12.924947

Построим **тепловую карту** *потребления электроэнергии* по дню недели и номеру недели года:



Отметим любопытный факт:

- какой-то явной закономерности в зависимости энергопотребления от дня недели в течении года обнаружить не удалось
- выделяется на фоне других по значению энергопотребления единственный день в году - вторник первой недели года. Это 1 января 2013
- в период с октября по апрель энергопотребление выше в течении всей недели, незначительно увеличиваясь в выходные дни

3.6. Подробнее исследуем зависимость потребления энергии от среднесуточных температур в течение года

3.6.1. Построим математическую модель зависимости энергопотребления от температуры наружного воздуха

Покажем *разброс среднего суточного потребления энергии всех домохозяйств* для каждого значения *среднесуточной температуры во всем диапазоне температур за год*. И покажем общий градиент.

Посчитаем среднесуточную температуру для каждых суток в выборке **data_weather_hourly_work_window**:

day	apparentTemperature	temperature
2013-01-01	2.160417	5.222500
2013-01-02	4.462083	6.717917
2013-01-03	10.277391	10.574348
2013-01-04	7.889583	9.641667
2013-01-05	8.045417	9.112500

Получим оценку максимального, минимального и медианного суточного потребления электроэнергии по всем домовладениям в этой выборке.

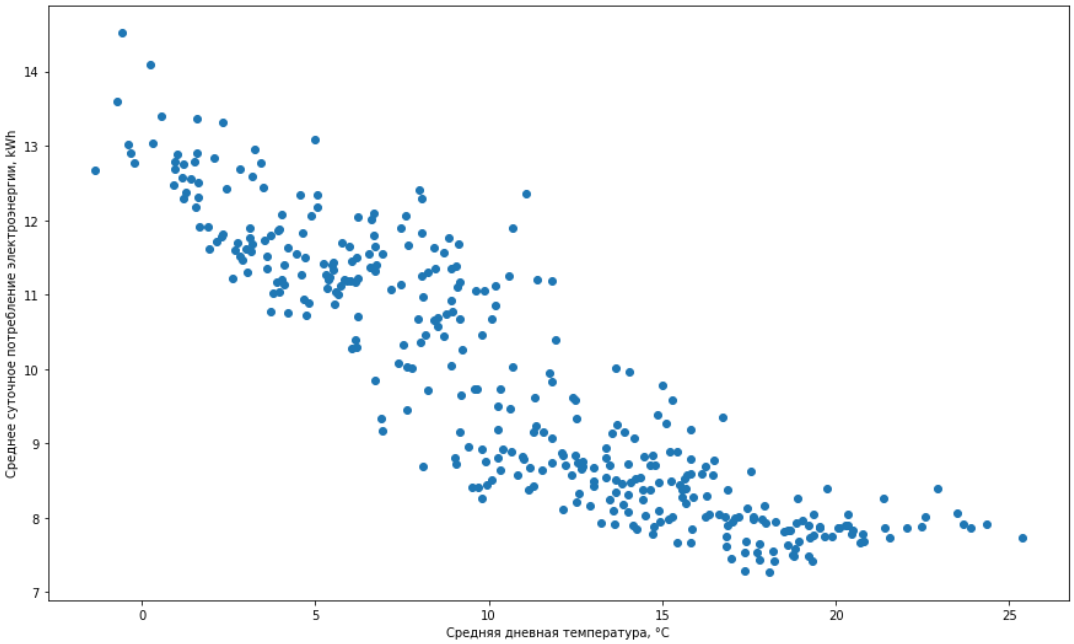
Максимальное суточное потребление одного домовладения в выборке составляет 136.56 КВтч, минимальное – 0 Квтч, медианное – 7.48 Квтч. ***Разрыв между максимумом и медианой колоссальный!*** Для дальнейшего анализа этого набора мы должны усреднить все значения потребленной энергии всех домовладений за каждые сутки.

Найдем *значения среднего суточного потребления* энергии всех домохозяйств в течении 2013 года и добавим в набор **data_daily_energy_mean** данные о среднесуточной температуре:

day	energy_sum	apparentTemperature	temperature
2013-01-01	11.417918	2.160417	5.222500
2013-01-02	11.312785	4.462083	6.717917
2013-01-03	11.251941	10.277391	10.574348
2013-01-04	11.052060	7.889583	9.641667
2013-01-05	11.683893	8.045417	9.112500

На ***графике покажем*** разброс *среднего суточного потребления* энергии всех домохозяйств для каждого значения среднесуточной температуры во всем диапазоне *температур за год*. И покажем общий градиент.

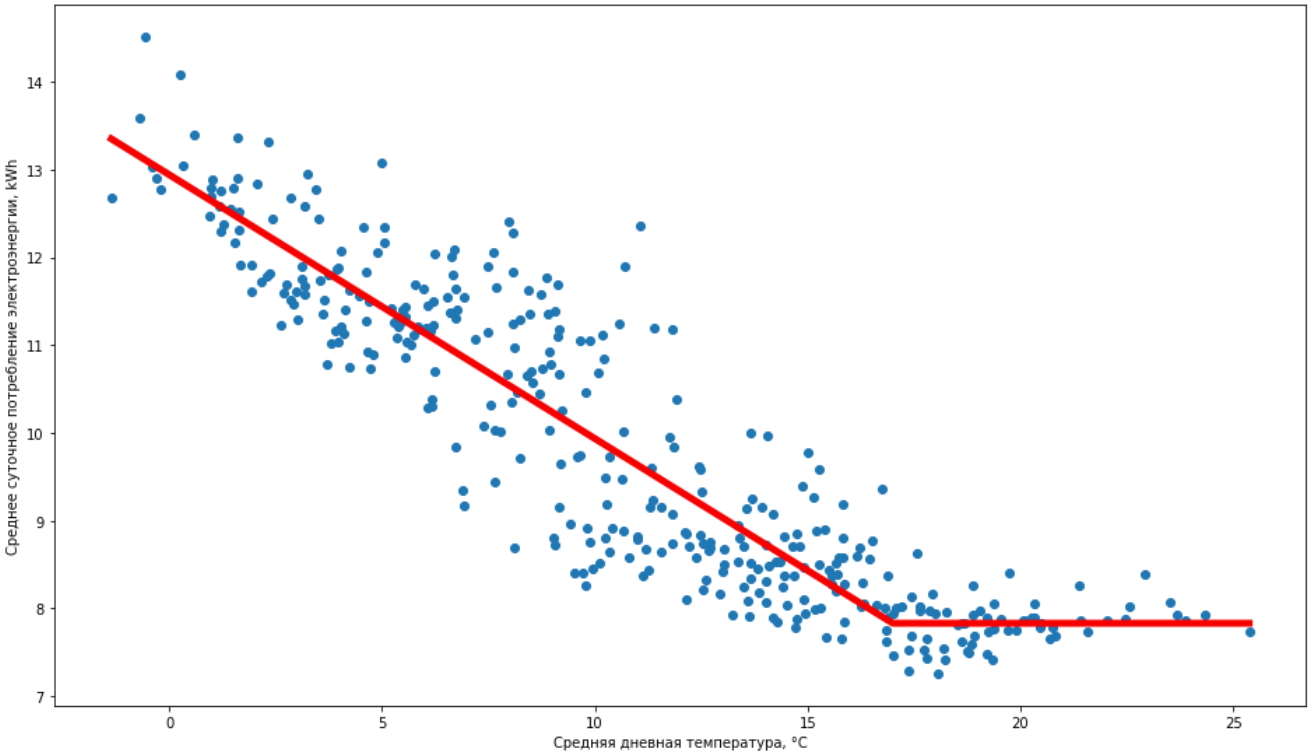
Разброс среднего суточного потребления энергии всех домохозяйств для каждого значения среднесуточной температуры во всем диапазоне температур за 2013 год



Покажем **общий градиент** *функции зависимости* энергопотребления от температуры наружного воздуха.

Из предыдущего *графика* видно, что эта функция имеет почти *линейный убывающий* характер в зоне *низких температур* и практически *не изменяется* при температурах *выше +14 (Цельсия)*. Отсюда следует, что градиент можно построить с помощью *функции кусочно-линейной регрессии*, использующей метод наименьших квадратов (**МНК**).

Кусочно-регрессионная модель зависимости энергопотребления от температуры наружного воздуха



У нас *получилась регрессионная модель* зависимости потребления энергии от наружной температуры в *зимний период* и прямая в *летний период*. Эта модель, называемая **PTG** (Power Temperature Gradient), часто используется при анализе энергоэффективности систем.

Более подробное описание модели можно *найти по ссылкам*:

- http://discovery.ucl.ac.uk/1473400/1/1-s2.0-S0378778815302905-main.pdf
- http://www.ibpsa.org/proceedings/BS2015/p2854.pdf

3.6.2. Найдем оценку R-квадрат (коэффициент детерминации; в Python - r2_score).

Она позволяет оценить меру зависимости одной случайной величины от множества других (в нашем случае потребленной энергии от температуры). Чем ближе значение R-квадрат к 1.0, тем эта зависимость более явная. Из результатов, полученных в предудущем п.3.7.1 можно сделать вывод, что если значение R-квадрат больше 70%, то часть домовладений использует электроэнергию для отопления в зимний период.

Коэффициент детерминации рассчитывается по формуле:

r2_score = r^2

где **r** - коэффициент корреляции.

Найдем корреляцию между величинами потребленной энергии и значениями температуры

Для нахождения *коэффициента корреляции* добавим следующие данные (здесь и далее *под температурой понимается температура комфорта*):

- **столбец 'xy'** - произведение средней температуры за день на суммарное потребление энергии за этот день;
- **столбец 'xx'** - квадрат дневной энергии;
- **столбец 'yy'** - квадрат средней температуры.

day	energy_sum	apparentTemperature	temperature	xy	xx	yy
2013-01-01	11.417918	2.160417	5.222500	24.667461	130.368856	4.667400
2013-01-02	11.312785	4.462083	6.717917	50.478589	127.979101	19.910188
2013-01-03	11.251941	10.277391	10.574348	115.640599	126.606172	105.624772
2013-01-04	11.052060	7.889583	9.641667	87.196145	122.148020	62.245525
2013-01-05	11.683893	8.045417	9.112500	94.001791	136.513366	64.728729

Подставляя найденные *числовые значения средних значений и среднеквадратических отклонений*, получаем искомый *коэффициент корреляции*: **-0.907475994929674**

Зная значение корреляции, можем рассчитать коэффициент детерминации:

r2_score = r^2 = 0.8235126813736018

Вывод: полученное *значение* **R-квадрат** показывает *высокую зависимость величин потребленной электроэнергии и значений измеренной температуры наружного воздуха в течение календарного года*. Следовательно, *наше предположение*, что часть домовладений использует электроэнергию для отопления в зимний период, *справедливо*.

3.6.3. Найдем долю домовладений, предположительно использующих электроэнергию для отопления.

Для этого найдем *отношение среднего потребления зимой и летом* для каждого домохозяйства. Будем считать, если зимнее потребление больше летнего *в 1.3 раза и выше*, то дом *отапливается электроэнергией*. Посчитаем долю таких домов в каждой из пяти групп потребителей.

Для оценки зимнего и летнего режимов энергопотребления будем использовать интервал в течение суток, в который любое энергопотребление, связанное с социальной и бытовой активностью сведено к минимуму - это ночное время суток (соответствует тарифному интервалу с минимальной нагрузкой на сеть - Т3).

Таким образом, энергопотребление в этом интервале суток должно быть примерно одинаковым в течении всего года, за исключением случаев использования электроэнергии для отопления дома.

Для вычислений будем использовать набор данных **data_hh_block_all_sun** (см. выше), уже содержащий информацию о ежесуточном потреблении энергии в тарифные интервалы **T1, T2, T3**.

Найдем *отношение* среднего потребления всех домовладений *в зимний период к летнему периоду*:

- Зима - 1255.39KWh,
- Лето - 895.48KWh
- Отношение потребления Зима/Лето - 1.40

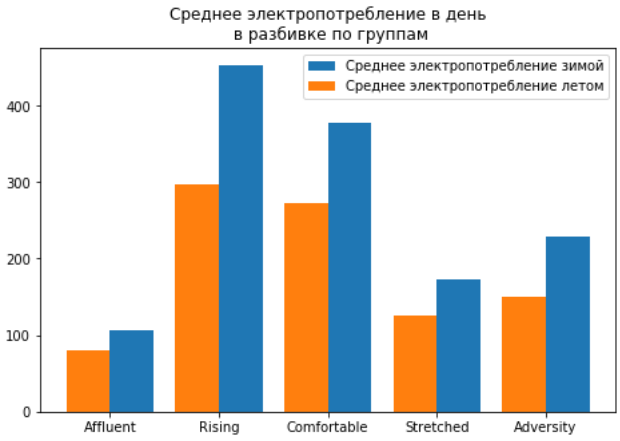
Нарисуем **диаграмму отношения** *среднесуточного энергопотребления Зима/Лето* всех домовладений:

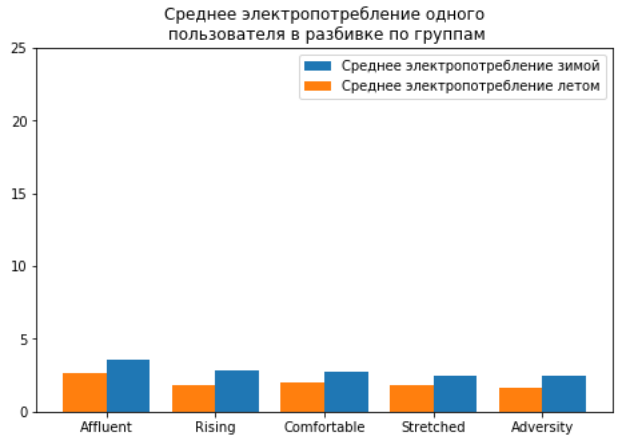
Теперь посчитаем *отношение среднего потребления Зима/Лето* в каждой из групп домовладений (**Acorn_grouped**) и определим, какие из домовладений *используют электроэнергию для отопления*:

- **'houses'**: число домовладений, отапливаемых в зимний сезон;
- **'winter_mean'**: значение среднего потребления электроэнергии всей группы в зимний сезон за период Т3;
- **'summer_mean'**: значение среднего потребления электроэнергии всей группы в летний сезон период Т3;
- **'total_houses'**: общее количество домовладений в группе.

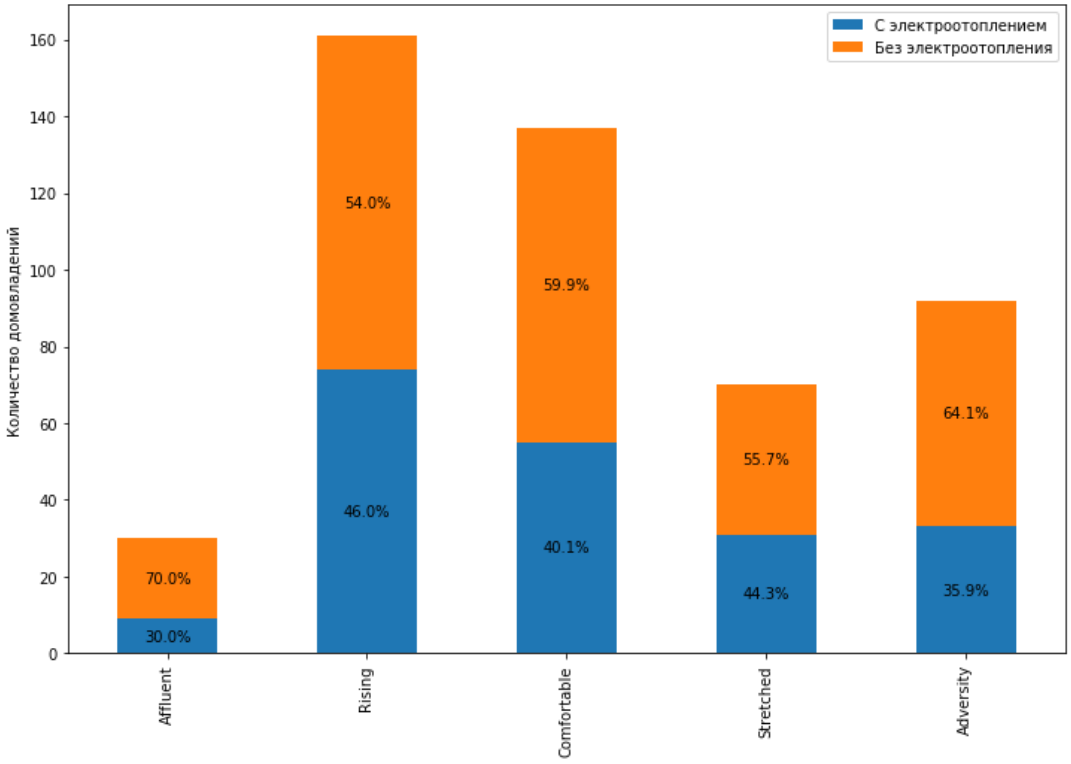
	houses	summer_mean	total_houses	winter_mean
Affluent	9.0	80.044424	30.0	106.507000
Rising	74.0	296.762989	161.0	452.748833
Comfortable	55.0	272.180696	137.0	377.361867
Stretched	31.0	125.144272	70.0	172.215089
Adversity	33.0	150.757696	92.0	228.523967

Проиллюстрируем полученные **результаты *графически***:





Доля домовладений, где использовалось электроотопление зимой в 2013г, в каждой из групп ACORN



Вывод:

- Судя по полученным результатам, нельзя сделать однозначных выводов о зависимости между уровнем благосостояния (по классификации ACORN) и использованием электроэнергии для отопления жилых помещений в зимний сезон
- Тем не менее можно констатировать, что имеется две группы, где число домовладений, использующих электроэнергию для отопления зимой, существенно больше чем в других группах. Это Affluent и Rising. Обе группы относятся к наиболее состоятельным.

4. Итоговые выводы

В данном проекте на выборке из показаний электросчетчиков 5566 домовладений Лондона (Великобритания), с ноября 2011 по февраль 2014, мы исследовали влияние различных факторов (погодных, социальных, экономических) на уровень потребления электрической энергии.

В первой и второй части работы мы нормировали выборку, отбросив наборы:

- с неполными данными:
  - малое число показаний счетчиков в течении суток
  - малое количество дней в году, когда проводились измерения потребления энергии
- с данными от домовладений, не имеющих общих критериев оценки факторов потребления электроэнергии с остальными домовладениями: общественные домовладения (не являющиеся жильем)
- определили как оптимальный период для исследования - один календарный год - 2013
- учитывая ограниченность вычислительных ресурсов, для исследования мы использовали только часть нормированной выборки, взяв из нее определенный % домовладений из каждой группы (Acorn)

В третьей части работы мы исследовали влияние на уровень энергопотребления следующих факторов:

- среднесуточной температуры
- выходных/праздничных и рабочих дней
- длины светового дня
- времени суток и дня недели в течении года
- выбранного тарифного плана

По результатам проведенного исследования мы пришли к следующим основным выводам:

- наибольшее влияние (до 70%) на динамику потребления электрической энергии оказывают два фактора
  - продолжительность светового дня
  - среднесуточная температура
- следующий по значимости фактор, оказывающий валяние (до 50%) на средние значение потребления энергии в расчете на одно домовладение - это принадлежность к определенной группе (Acorn), определяемая уровнем благосостояния. Чем этот уровень выше - тем больше энергопотребление.
- значительного влияния (больше 2-10%) остальных факторов на динамику энергопотребления установить не удалось

При исследовании зависимости суммарного потребления энергии в течении года от времени суток (от дня недели) мы использовали качественную оценку динамики энергопотребления с помощью тепловых карт. Тепловые карты наглядно иллюстрируют нам следующие факты:

- самое большое потребление электроэнергии в течении года происходит в вечерние часы с 17:00 24:00

- на втором месте по потреблению дневные часы с 9:00 до 17:00
- пиковые значения приходятся на часы с 18:30 до 22:00 в период с конца октября по начало апреля
- в период с октября по апрель энергопотребление выше в течении всей недели, и незначительно увеличивается в выходные дни

Дополнительно мы подробно исследовали зависимость потребления энергии от среднесуточных температур наружного воздуха в течении года:

- У нас получилась регрессионная модель зависимости потребления энергии от наружной температуры в зимний период и прямая в летний период;
- Проведя расчет корреляции и коэффициента детерминации (его значение выше 0.8) мы доказали высокую зависимость этих величин;
- Основываясь на этом факте мы сделали предположение и доказали факт использования электроэнергии для отопления части домовладений в зимний период;
- Мы посчитали долю домовладений в каждой из групп, использующих зимой электроэнергию для отопления. Доля таких домовладения составила от 48% (в более обеспеченных группах) до 33% (в мало обеспеченных группах).