In [2]:
```
import numpy as nü
from matplotlib import pyplot as plt
import pandas as pd
```

In [3]:
```
metadata = pd.read_csv("ashrae3-building_metadata.csv")
ashrae = pd.read_csv("ashrae3-train.csv")
weather = pd.read_csv("ashrae3-weather_train.csv")
```

In [ ]:
```
```

In [4]:
```
metadata.head()
```

Out[4]:

	site_id	building_id	primary_use	square_feet	year_built	floor_count
0	0	0	Education	7432	2008.0	NaN
1	0	1	Education	2720	2004.0	NaN
2	0	2	Education	5376	1991.0	NaN
3	0	3	Education	23685	2002.0	NaN
4	0	4	Education	116607	1975.0	NaN

In [5]:
```
ashrae.head()
```

Out[5]:

	building_id	meter	timestamp	meter_reading
0	0	0	2016-01-01 00:00:00	0.0
1	1	0	2016-01-01 00:00:00	0.0
2	2	0	2016-01-01 00:00:00	0.0
3	3	0	2016-01-01 00:00:00	0.0
4	4	0	2016-01-01 00:00:00	0.0

In [6]:
```
weather.head()
```

Out[6]:

	site_id	timestamp	air_temperature	cloud_coverage	dew_temperature	precip_depth_1_hr	sea_level_pressure	wind_direction	wind_sp
0	0	2016-01-01 00:00:00	25.0	6.0	20.0	NaN	1019.7	0.0	
1	0	2016-01-01 01:00:00	24.4	NaN	21.1	-1.0	1020.2	70.0	
2	0	2016-01-01 02:00:00	22.8	2.0	21.1	0.0	1020.2	0.0	
3	0	2016-01-01 03:00:00	21.1	2.0	20.6	0.0	1020.1	0.0	
4	0	2016-01-01 04:00:00	20.0	2.0	20.0	-1.0	1020.0	250.0	

In [7]:
```
#2: hängt der Stromverbauch im engen Zusammenhang mit der Größe des Gebäudes?
#   Hängt der Stromverbrauch im Zusammenhang mit dem Nutzen zusammen ?
#   Wie ist der Stromvebrauch verteilt ?
#   Hängt die Lufttemperatur mit dem Stromvebrauch zusammen ?
```

In [9]:
```
#3
#Meine Vermutung liegt nahe, dass floor_count zu viele null-werte beinhaltet.
#Und ich empfinde auch floor_count um meine Forschungsfrage zu beantworten nicht präzise.
#Ein Gebäude kann viele Etagen haben aber dies bedeutet nicht dass dieses Gebäude hoch oder breit is
t.
#Die Etagenanzahl gibt nicht wirklich auskunft über die Höhe und Breite der Gebäude viel eher der Sq
uare-Feet
#Aus diesem Grund habe ich die floor-count gelöscht.
#Ebenso empfinde ich dass das Gebaudejahr nicht relevant fuer den Stromvebrauch ist.
#Klar verbauchen ältere Gebäude mehr Strom aufgrund der veralteten Technolgie aber für die Vorhersag
e ist
#dies nicht entscheidend, daher lösche ich diese Spalte

print(metadata['floor_count'].isna().sum())
print(metadata['square_feet'].isna().sum())
metadataclean = metadata.drop(['floor_count','year_built'], axis=1)
metadataclean
```

```
1094
0
```

Out[9]:

	site_id	building_id	primary_use	square_feet
0	0	0	Education	7432
1	0	1	Education	2720
2	0	2	Education	5376
3	0	3	Education	23685
4	0	4	Education	116607
...	...	...	...	...
1444	15	1444	Entertainment/public assembly	19619
1445	15	1445	Education	4298
1446	15	1446	Entertainment/public assembly	11265
1447	15	1447	Lodging/residential	29775
1448	15	1448	Office	92271

1449 rows × 4 columns

In [30]:
```
#3
#Meter reading gibt den Stromvebrauch an, viele Werte beinhalten 0 Werte.
#Dies liegt daran, weil sie an dem Tag errichtet worden sind.
#Man muss bedenken, dass diese Werte die Vorhersage stark beeinflussen.
#Meine Entscheidung ist, dass ich diese Daten selektieren werde und jeweils nur Werte größer
print(len(ashrae[ashrae['meter_reading'] == 0]))
print(len(ashrae[ashrae['meter_reading'] != 0]))

p = len(ashrae[ashrae['meter_reading'] == 0]) / len(ashrae[ashrae['meter_reading'] != 0])

ashrae_clean = ashrae[ashrae['meter_reading'] != 0]

ashrae_clean
```

```
1873976
18342124
```

Out[30]:

	building_id	meter	timestamp	meter_reading
45	46	0	2016-01-01 00:00:00	53.2397
72	74	0	2016-01-01 00:00:00	43.0013
91	93	0	2016-01-01 00:00:00	52.4206
103	105	0	2016-01-01 00:00:00	23.3036
104	106	0	2016-01-01 00:00:00	0.3746
...	...	...	...	...
20216094	1443	0	2016-12-31 23:00:00	64.9500
20216095	1444	0	2016-12-31 23:00:00	8.7500
20216096	1445	0	2016-12-31 23:00:00	4.8250
20216098	1447	0	2016-12-31 23:00:00	159.5750
20216099	1448	0	2016-12-31 23:00:00	2.8500

18342124 rows × 4 columns

In [ ]:
```
#ich hoffe, dass dies reicht für mind 8 punkte :)
```