...alysis 1 - Spark DF

610 Analysis 1 - Spark DF

Settings

- Check the active interpreters: md (needs to be on the top), spark2, sh, shUser.
- File for project:

> FILE='prep_household_data_1min_singleindex.csv'

Approach

- Download the data to local device.
- Manipulate the data with Pandas (Pycharm).
- Upload the data to the gateway.

Data Manipulation with Pandas

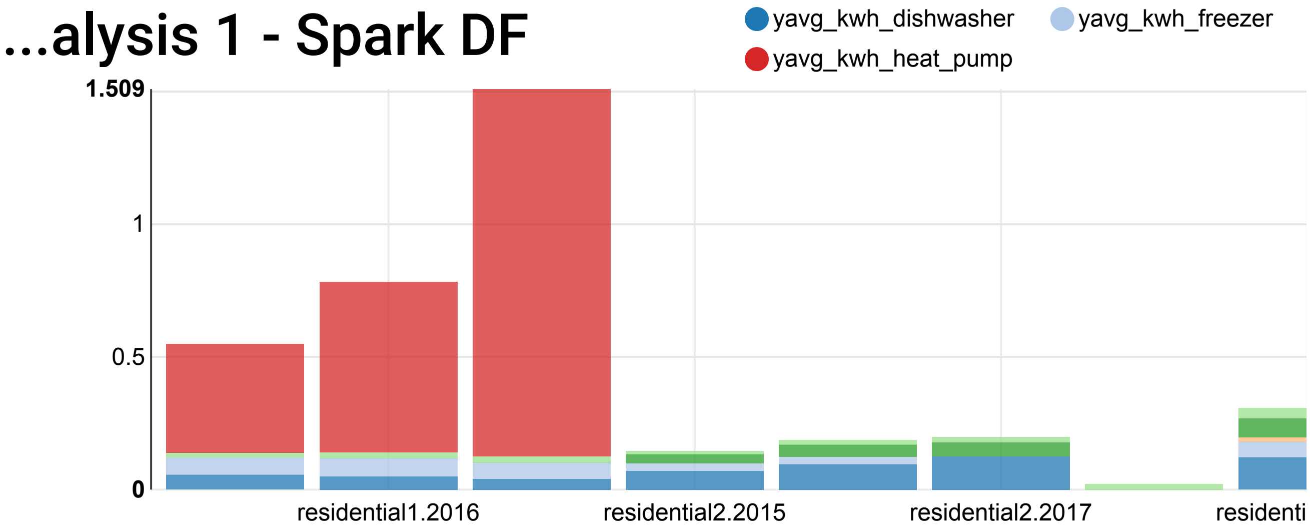
> data_preprocessing.py

```
total 1.8G
-rw-rw-r-- 1 bd01 bd01 119M Feb  5 19:02 prep_household_data_15min_singleindex.csv
-rw-rw-r-- 1 bd01 bd01 1.6G Feb  5 19:00 prep_household_data_1min_singleindex.csv
-rw-rw-r-- 1 bd01 bd01  34M Feb  5 18:56 prep_household_data_60min_singleindex.csv
----hdfs----
Found 6 items
drwx------   - zeppelin hdfs          0 2021-02-06 06:00 /user/zeppelin/.Trash
drwxr-xr-x   - zeppelin hdfs          0 2021-02-07 14:23 /user/zeppelin/conf
drwxr-xr-x   - zeppelin hdfs          0 2021-02-07 16:50 /user/zeppelin/notebook
-rw-r--r--   1 zeppelin hdfs  124520554 2021-02-05 20:36 /user/zeppelin/prep_household_data_15min_singlein
dex.csv
-rw-r--r--   1 zeppelin hdfs 1712423587 2021-02-06 11:26 /user/zeppelin/prep_household_data_1min_singleind
ex.csv
-rw-r--r--   1 zeppelin hdfs   34626321 2021-02-05 20:35 /user/zeppelin/prep_household_data_60min_singlein
dex.csv
```

Load into a Spark dataframe

Analysis: Which devices consume the majority of households

energy (average kWh)?



First analysis shows a large variability among the largest energy consumer household devices. Whereas heat pumps consume by far the largest amount of households energy, electric vehicles probably follow second. The energy consumation of other factors such as dishwasher and freezer show huge variation across households. The usage of these devices depends on individual behaviour. Unfortunately, the data does not provide an indicator for household size or living area.