490 Project - Python DataFrames

December 1, 2020

In [39]:
```
import csv
from pyspark.sql import SparkSession
import pyspark.sql.functions as f
# User Defined Functions
from pyspark.sql.functions import udf
from pyspark.sql.types import DoubleType
# Stats
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.stat import Correlation
from math import sqrt

import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
```

In [40]:
```
# Creates a new Spark session w/in Python
spark = SparkSession.builder.appName("Final Project").getOrCreate()
```

1 The Data

The dataset our group chose was based on hourly energy demand for 5 cities in Spain (https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather). The goal of this Jupyter Notebook is to explore the data to look for important relationships in the data

This will be done using PySpark as we have learned in class and matplotlib for visualiztion

2 Data Setup

Importing the data

In [41]:
```
# Reads the local csv stored on my computer, it does have a header
energy_df = spark.read.csv("C:\\Users\\Wes\\Desktop\\CS 490\\energy_dataset.csv", hea
weather_df = spark.read.csv("C:\\Users\\Wes\\Desktop\\CS 490\\weather_features.csv", h
# Creates a temporary view from the dataframe we read from the file
# This is how we can read SQL from it
```

```
energy_df.createOrReplaceTempView("Energy")
weather_df.createOrReplaceTempView("Weather")
```

Creating the Combined dataframe

```
In [42]: joined_df = energy_df.join(weather_df, energy_df.time == weather_df.dt_iso)
         joined_df.show(1, vertical=True)
```

```
-RECORD 0-----------------------------------------------------------
 time                                         | 2015-01-01 00:00:00
 generation_biomass                           | 447.0
 generation_fossil_brown_coal_lignite         | 329.0
 generation_fossil_coal_derived_gas           | 0.0
 generation_fossil_gas                        | 4844.0
 generation_fossil_hard_coal                  | 4821.0
 generation_fossil_oil                        | 162.0
 generation_fossil_oil_shale                  | 0.0
 generation_fossil_peat                       | 0.0
 generation_geothermal                        | 0.0
 generation_hydro_pumped_storage_aggregated   | null
 generation_hydro_pumped_storage_consumption  | 863.0
 generation_hydro_run_of_river_and_poundage   | 1051.0
 generation_hydro_water_reservoir             | 1899.0
 generation_marine                            | 0.0
 generation_nuclear                           | 7096.0
 generation_other                             | 43.0
 generation_other_renewable                   | 73.0
 generation_solar                             | 49.0
 generation_waste                             | 196.0
 generation_wind_offshore                     | 0.0
 generation_wind_onshore                      | 6378.0
 forecast_solar_day_ahead                     | 17.0
 forecast_wind_offshore_eday_ahead            | null
 forecast_wind_onshore_day_ahead              | 6436.0
 total_load_forecast                          | 26118.0
 total_load_actual                            | 25385.0
 price_day_ahead                              | 50.1
 price_actual                                 | 65.41
 dt_iso                                       | 2015-01-01 00:00:00
 city_name                                    | Valencia
 temp                                         | 270.475
 temp_min                                     | 270.475
 temp_max                                     | 270.475
 pressure                                     | 1001
 humidity                                     | 77
 wind_speed                                   | 1
 wind_deg                                     | 62
 rain_1h                                      | 0.0
```

```
 rain_3h                                  | 0.0
 snow_3h                                  | 0.0
 clouds_all                               | 0
 weather_id                               | 800
 weather_main                             | clear
 weather_description                      | sky is clear
 weather_icon                             | 01n
only showing top 1 row
```

Creating columns for Farenheight temperatures and filling na values

```
In [43]: # Convert Kelvin temps to Farhenheight
         k_to_f_udf = udf(lambda kelvin: (float(kelvin) - 273.15) * (9/5) + 32, DoubleType())
         joined_df = joined_df.withColumn("temp_f", k_to_f_udf(joined_df.temp))
         joined_df = joined_df.withColumn("temp_min_f", k_to_f_udf(joined_df.temp_min))
         joined_df = joined_df.withColumn("temp_max_f", k_to_f_udf(joined_df.temp_max))
         joined_df.na.fill(0)

Out[43]: DataFrame[time: string, generation_biomass: double, generation_fossil_brown_coal_lign
```

3 Data Exploration

3.1 By City

Based on the 5 cities within the dataset I wanted to see how the data varied between them

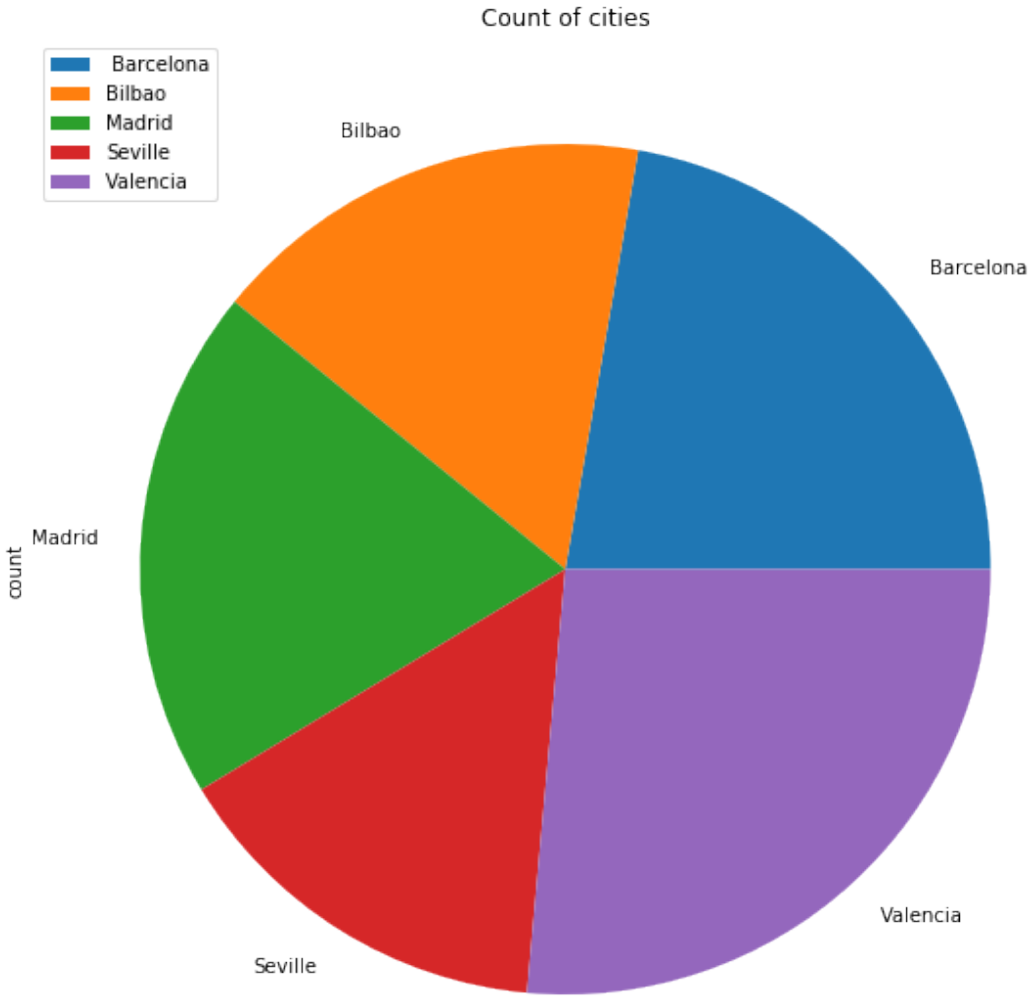
```
In [44]: test_agg = joined_df.groupBy("city_name", "weather_main")\
             .agg({"time" : "count", "price_actual" : "avg"})\
             .withColumnRenamed('count(time)', 'count')\
             .withColumnRenamed('avg(price_actual)', 'avg_price')\
             .filter("count > 100")\
             .orderBy("city_name", "weather_main")
         test_agg.show(100, False)
```

```
+----------+------------+------------------+-----+
|city_name |weather_main|avg_price         |count|
+----------+------------+------------------+-----+
| Barcelona|clear       |59.30181183960987 |14764|
| Barcelona|clouds      |57.059017502482575|16112|
| Barcelona|drizzle     |56.17348837209302 |301  |
| Barcelona|mist        |55.9887133182844  |443  |
| Barcelona|rain        |55.62026551226556 |3465 |
| Barcelona|thunderstorm|62.42364820846908 |307  |
|Bilbao    |clear       |60.56761825922418 |8456 |
|Bilbao    |clouds      |57.37178233815964 |16714|
|Bilbao    |drizzle     |60.67769072164944 |485  |
|Bilbao    |fog         |62.91078397212544 |1148 |
```

```
|Bilbao    |mist        |59.97895819508959 |1507 |
|Bilbao    |rain        |54.971499514495804|7209 |
|Bilbao    |snow        |55.17807228915661 |166  |
|Bilbao    |thunderstorm|58.42384615384614 |208  |
|Madrid    |clear       |58.671028877320325|20362|
|Madrid    |clouds      |56.2445749178017  |10645|
|Madrid    |drizzle     |57.66627943485085 |637  |
|Madrid    |fog         |64.63083333333327 |708  |
|Madrid    |mist        |62.92892324093817 |938  |
|Madrid    |rain        |56.281083929243515|2657 |
|Madrid    |thunderstorm|63.46621621621619 |222  |
|Seville   |clear       |58.691902238426174|23588|
|Seville   |clouds      |56.31560800552104 |7245 |
|Seville   |drizzle     |49.820041152263364|243  |
|Seville   |dust        |58.34052173913042 |345  |
|Seville   |fog         |56.43466292134829 |534  |
|Seville   |haze        |53.48614942528738 |348  |
|Seville   |mist        |59.37509638554222 |830  |
|Seville   |rain        |55.29623897707229 |2268 |
|Seville   |thunderstorm|57.56370370370372 |135  |
|Valencia  |clear       |57.33447268515524 |15541|
|Valencia  |clouds      |58.23793175005739 |17348|
|Valencia  |mist        |56.55921052631578 |190  |
|Valencia  |rain        |58.80600779510035 |1796 |
|Valencia  |thunderstorm|63.03674556213019 |169  |
+----------+------------+------------------+-----+
```

In [45]:
```
graph_df = test_agg.select('*').toPandas()
plt.rcParams["figure.figsize"] = 18,10
graph_df.groupby(['city_name']).mean().plot(kind='pie',y='count', title='Count of citi
```

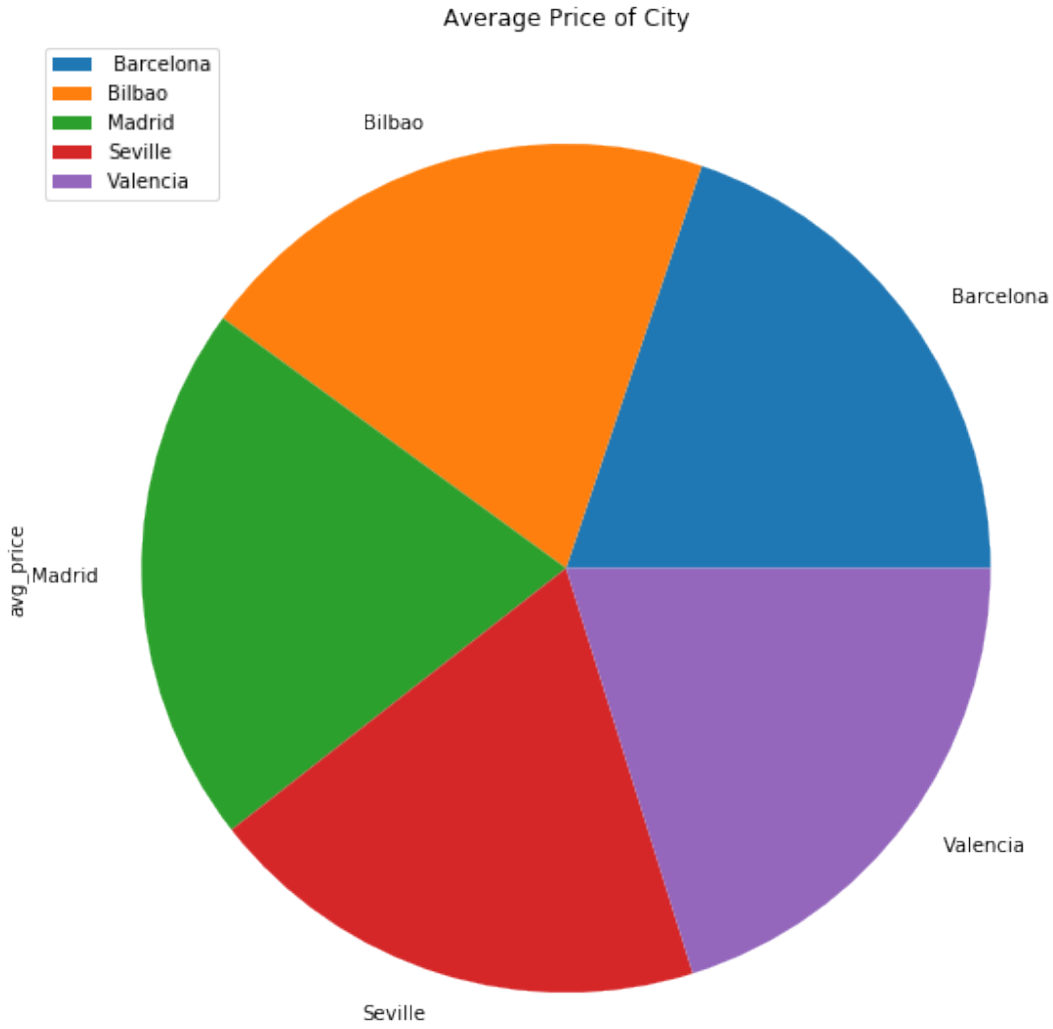
Out[45]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x146569f0>
```



The representation of data between the cities within the dataset is very similar, with Valencia being the most represented and Seville being the least represented

```
In [46]: graph_df.groupby(['city_name']).mean().plot(kind='pie',y='avg_price', title='Average
```

```
Out[46]: <matplotlib.axes._subplots.AxesSubplot at 0x153b27f0>
```



The average price between each city is nearly identical as well. This is something we had hoped to see, as it means the pricing is fair between each city and no one customer is paying more than another just for living in a different city

4 Correlation

```
In [47]: def correlation_matrix(df, corr_columns, method='pearson'):
             vector_col = "corr_features"
             assembler = VectorAssembler(inputCols=corr_columns, outputCol=vector_col)
             df_vector = assembler.setHandleInvalid("keep").transform(df).select(vector_col)
             np.nan_to_num(df_vector)
             matrix = Correlation.corr(df_vector, vector_col, method)
```

```
            result = matrix.collect()[0]["pearson({})".format(vector_col)].values
            return pd.DataFrame(result.reshape(-1, len(corr_columns)), columns=corr_columns, i
```

The above function takes in a (Pyspark-style) dataframe and a list of columns to create a Correlation matrix of the values

4.1 By Wind Speed

Does the price depend on the wind speed?

```
In [48]: test_agg = joined_df.groupBy("wind_speed")\
             .agg({"time" : "count", "price_actual" : "avg"})\
             .withColumnRenamed('count(time)', 'count')\
             .withColumnRenamed('avg(price_actual)', 'avg_price')\
             .filter("count > 100")\
             .sort("wind_speed")
         test_agg.show(100, False)
```

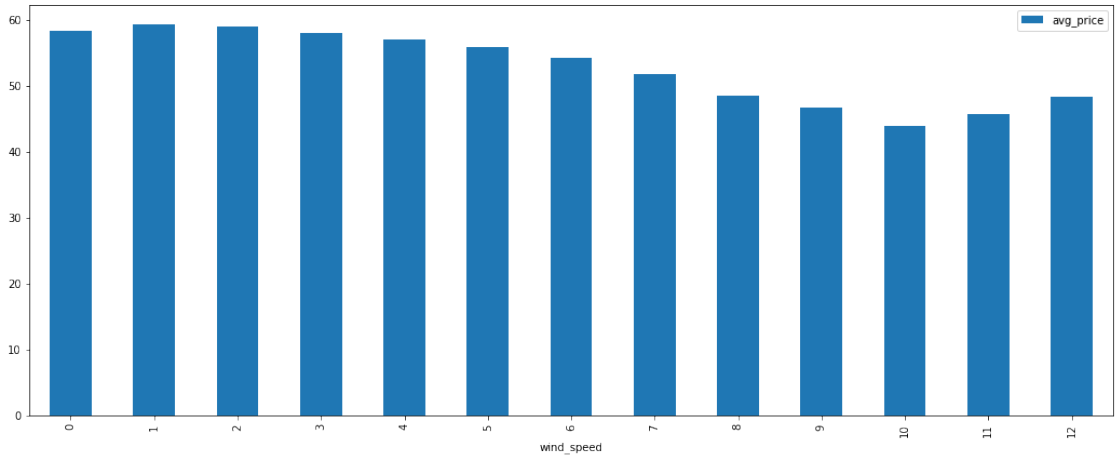
```
+----------+------------------+-----+
|wind_speed|avg_price         |count|
+----------+------------------+-----+
|0         |58.48113213667815 |18496|
|1         |59.43090143599589 |55223|
|2         |59.01943278830858 |34555|
|3         |58.0480008785592  |25041|
|4         |57.06761207885105 |18313|
|5         |55.98622442865706 |11683|
|6         |54.275993523697394|6794 |
|7         |51.87357766604919 |3779 |
|8         |48.61905970850964 |2127 |
|9         |46.72045927209707 |1154 |
|10        |43.998986866791746|533  |
|11        |45.793608562691134|327  |
|12        |48.36364532019704 |203  |
+----------+------------------+-----+
```

```
In [49]: correlation_matrix(test_agg, test_agg.columns)
```

```
Out[49]:             wind_speed  avg_price     count
         wind_speed    1.000000  -0.934583 -0.814867
         avg_price    -0.934583   1.000000  0.807062
         count        -0.814867   0.807062  1.000000
```

```
In [50]: graph_df = test_agg.select('*').toPandas()
         plt.rcParams["figure.figsize"] = 18,7
         graph_df.plot.bar(x='wind_speed',
                           y='avg_price')
```

```
Out[50]: <matplotlib.axes._subplots.AxesSubplot at 0x153f19b0>
```



There seems to be a strong negative correlation between wind speed and average price, although as the wind speeds get higher, there is less and less of a count of these represented in the data, so higher values aren't as trustworthy as the lower ones.

4.2 By Month and Day

```
In [51]: test_agg = joined_df.groupBy(f.month('time'), f.dayofmonth('time'))\
             .agg({"time" : "count", "price_actual" : "avg"})\
             .withColumnRenamed('dayofmonth(time)', 'day_of_month')\
             .withColumnRenamed('month(time)', 'month')\
             .withColumnRenamed('count(time)', 'count')\
             .withColumnRenamed('avg(price_actual)', 'avg_price')\
             .orderBy('month', 'day_of_month')\
             .filter("count > 100")
         test_agg.show(10,False)
```

```
+-----+------------+------------------+-----+
|month|day_of_month|avg_price         |count|
+-----+------------+------------------+-----+
|1    |1           |46.19534161490682 |483  |
|1    |2           |56.790477178423245|482  |
|1    |3           |54.05941414141414 |495  |
|1    |4           |55.018605577689236|502  |
|1    |5           |60.3349007936508  |504  |
|1    |6           |55.52820512820514 |507  |
|1    |7           |59.728073217726376|519  |
|1    |8           |65.08919087136927 |482  |
|1    |9           |64.61051792828685 |502  |
|1    |10          |56.60917647058824 |510  |
+-----+------------+------------------+-----+
```

```
only showing top 10 rows
```

```
In [52]: correlation_matrix(test_agg, test_agg.columns)
```

```
Out[52]:                  month  day_of_month  avg_price     count
         month         1.000000      0.006443   0.595330  0.081220
         day_of_month  0.006443      1.000000   0.045038 -0.073601
         avg_price     0.595330      0.045038   1.000000  0.259623
         count         0.081220     -0.073601   0.259623  1.000000
```

According to the correlations, the average price seems to depend much more on what month it is than the day of the month. This tracks as we expect energy costs to vary throughout the year as temperatures change, as opposed to costs varying wildly over the course of a month

4.3 Temperature, price, and actual load by year and month

Ignoring the day of the month and looking at how temmperature, prices and actual load depend by year and month

```
In [53]: test_agg = joined_df.groupBy(f.year('time'), f.month('time'))\
             .agg({"price_actual" : "avg", \
                   "total_load_actual" : "avg",\
                   "temp_f": "avg",\
                   "temp_min_f": "avg",\
                   "temp_max_f": "avg"})\
             .withColumnRenamed('year(time)', 'year')\
             .withColumnRenamed('month(time)', 'month')\
             .withColumnRenamed('avg(price_actual)', 'avg_price')\
             .withColumnRenamed('avg(total_load_actual)', 'avg_total_load_actual')\
             .orderBy('year','month')
```

```
In [54]: correlation_matrix(test_agg, test_agg.columns)
```

```
Out[54]:                            year     month  avg(temp_f)  avg(temp_max_f)  \
         year                   1.000000  0.000000    -0.024369        -0.061868
         month                  0.000000  1.000000     0.276436         0.286410
         avg(temp_f)           -0.024369  0.276436     1.000000         0.989586
         avg(temp_max_f)       -0.061868  0.286410     0.989586         1.000000
         avg(temp_min_f)        0.012814  0.260331     0.989929         0.959802
         avg_total_load_actual  0.304222 -0.210858    -0.104184        -0.124596
         avg_price              0.201002  0.389448     0.023399        -0.008685

                            avg(temp_min_f)  avg_total_load_actual  avg_price
         year                      0.012814               0.304222   0.201002
         month                     0.260331              -0.210858   0.389448
         avg(temp_f)               0.989929              -0.104184   0.023399
         avg(temp_max_f)           0.959802              -0.124596  -0.008685
```

```
avg(temp_min_f)              1.000000             -0.084801   0.047780
avg_total_load_actual       -0.084801              1.000000   0.283745
avg_price                    0.047780              0.283745   1.000000
```

At this level of time granularity there are no strong correlations (not counting temperature relates to itself). Month and Average Price are the two strongest correlation present in this relation

4.4 Generation_ by Year

The energy dataset contains many columns concerning the generation of power (measured in MegaWatts). It's important to look at how these relate to the price and total load. This will also do us an important service of telling us if there are any unimportant columns that we can ignore

```
In [55]: test_agg = joined_df.groupBy(f.year('time'))\
             .agg({"price_actual" : "avg", \
                   "total_load_actual" : "avg",\
                   "generation_biomass" : "avg",\
                   "generation_fossil_brown_coal_lignite" : "avg",\
                   "generation_fossil_coal_derived_gas" : "avg",\
                   "generation_fossil_gas" : "avg",\
                   "generation_fossil_hard_coal" : "avg",\
                   "generation_fossil_oil" : "avg",\
                   "generation_fossil_oil_shale" : "avg",\
                   "generation_fossil_peat" : "avg",\
                   "generation_geothermal" : "avg",\
                   "generation_hydro_pumped_storage_aggregated" : "avg",\
                   "generation_hydro_pumped_storage_consumption" : "avg",\
                   "generation_hydro_run_of_river_and_poundage" : "avg",\
                   "generation_hydro_water_reservoir" : "avg",\
                   "generation_marine" : "avg",\
                   "generation_nuclear" : "avg",\
                   "generation_other" : "avg",\
                   "generation_other_renewable" : "avg",\
                   "generation_solar" : "avg",\
                   "generation_waste" : "avg",\
                   "generation_wind_offshore" : "avg",\
                   "generation_wind_onshore" : "avg",\
                  })\
             .withColumnRenamed('year(time)', 'year')\
             .orderBy('year')

In [56]: correlation_matrix(test_agg, test_agg.columns)

Out[56]:                                                    year  \
         year                                           1.000000
         avg(generation_fossil_gas)                     0.707407
         avg(generation_fossil_brown_coal_lignite)     -0.443560
         avg(generation_hydro_pumped_storage_aggregated)      NaN
         avg(generation_fossil_peat)                          NaN
```

```
avg(generation_wind_offshore)                          NaN
avg(generation_hydro_water_reservoir)            -0.112402
avg(total_load_actual)                            0.988030
avg(generation_solar)                            -0.376027
avg(generation_geothermal)                             NaN
avg(generation_other)                            -0.863683
avg(price_actual)                                 0.325889
avg(generation_other_renewable)                   0.974178
avg(generation_nuclear)                          -0.508627
avg(generation_hydro_pumped_storage_consumption) -0.870629
avg(generation_biomass)                          -0.863918
avg(generation_marine)                                 NaN
avg(generation_waste)                             0.950379
avg(generation_fossil_hard_coal)                 -0.701993
avg(generation_fossil_coal_derived_gas)                NaN
avg(generation_hydro_run_of_river_and_poundage)   0.522804
avg(generation_fossil_oil_shale)                       NaN
avg(generation_wind_onshore)                      0.514543
avg(generation_fossil_oil)                       -0.796692

                                                 avg(generation_fossil_gas)  \
year                                                               0.707407
avg(generation_fossil_gas)                                         1.000000
avg(generation_fossil_brown_coal_lignite)                          0.296033
avg(generation_hydro_pumped_storage_aggregated)                         NaN
avg(generation_fossil_peat)                                             NaN
avg(generation_wind_offshore)                                           NaN
avg(generation_hydro_water_reservoir)                             -0.750713
avg(total_load_actual)                                             0.772630
avg(generation_solar)                                              0.386652
avg(generation_geothermal)                                              NaN
avg(generation_other)                                             -0.823875
avg(price_actual)                                                  0.379506
avg(generation_other_renewable)                                    0.831806
avg(generation_nuclear)                                           -0.028780
avg(generation_hydro_pumped_storage_consumption)                  -0.819886
avg(generation_biomass)                                           -0.678732
avg(generation_marine)                                                  NaN
avg(generation_waste)                                              0.855207
avg(generation_fossil_hard_coal)                                  -0.241981
avg(generation_fossil_coal_derived_gas)                                 NaN
avg(generation_hydro_run_of_river_and_poundage)                   -0.230797
avg(generation_fossil_oil_shale)                                        NaN
avg(generation_wind_onshore)                                      -0.067438
avg(generation_fossil_oil)                                        -0.424401

                                                 avg(generation_fossil_brown_coal_li
year                                                                          -0.4
```

avg(generation_fossil_gas)	0.1
avg(generation_fossil_brown_coal_lignite)	1.0
avg(generation_hydro_pumped_storage_aggregated)	
avg(generation_fossil_peat)	
avg(generation_wind_offshore)	
avg(generation_hydro_water_reservoir)	-0.8
avg(total_load_actual)	-0.3
avg(generation_solar)	0.9
avg(generation_geothermal)	
avg(generation_other)	0.1
avg(price_actual)	0.1
avg(generation_other_renewable)	-0.1
avg(generation_nuclear)	0.4
avg(generation_hydro_pumped_storage_consumption)	0.0
avg(generation_biomass)	0.4
avg(generation_marine)	
avg(generation_waste)	-0.1
avg(generation_fossil_hard_coal)	0.7
avg(generation_fossil_coal_derived_gas)	
avg(generation_hydro_run_of_river_and_poundage)	-0.9
avg(generation_fossil_oil_shale)	
avg(generation_wind_onshore)	-0.6
avg(generation_fossil_oil)	0.6

	avg(generation_hydro_pumped_storage_
year	
avg(generation_fossil_gas)	
avg(generation_fossil_brown_coal_lignite)	
avg(generation_hydro_pumped_storage_aggregated)	
avg(generation_fossil_peat)	
avg(generation_wind_offshore)	
avg(generation_hydro_water_reservoir)	
avg(total_load_actual)	
avg(generation_solar)	
avg(generation_geothermal)	
avg(generation_other)	
avg(price_actual)	
avg(generation_other_renewable)	
avg(generation_nuclear)	
avg(generation_hydro_pumped_storage_consumption)	
avg(generation_biomass)	
avg(generation_marine)	
avg(generation_waste)	
avg(generation_fossil_hard_coal)	
avg(generation_fossil_coal_derived_gas)	
avg(generation_hydro_run_of_river_and_poundage)	
avg(generation_fossil_oil_shale)	
avg(generation_wind_onshore)	

```
avg(generation_fossil_oil)

                                                    avg(generation_fossil_peat)  \
year                                                                        NaN
avg(generation_fossil_gas)                                                  NaN
avg(generation_fossil_brown_coal_lignite)                                   NaN
avg(generation_hydro_pumped_storage_aggregated)                             NaN
avg(generation_fossil_peat)                                                 1.0
avg(generation_wind_offshore)                                               NaN
avg(generation_hydro_water_reservoir)                                       NaN
avg(total_load_actual)                                                      NaN
avg(generation_solar)                                                       NaN
avg(generation_geothermal)                                                  NaN
avg(generation_other)                                                       NaN
avg(price_actual)                                                           NaN
avg(generation_other_renewable)                                             NaN
avg(generation_nuclear)                                                     NaN
avg(generation_hydro_pumped_storage_consumption)                            NaN
avg(generation_biomass)                                                     NaN
avg(generation_marine)                                                      NaN
avg(generation_waste)                                                       NaN
avg(generation_fossil_hard_coal)                                            NaN
avg(generation_fossil_coal_derived_gas)                                     NaN
avg(generation_hydro_run_of_river_and_poundage)                             NaN
avg(generation_fossil_oil_shale)                                            NaN
avg(generation_wind_onshore)                                                NaN
avg(generation_fossil_oil)                                                  NaN

                                                    avg(generation_wind_offshore)  \
year                                                                          NaN
avg(generation_fossil_gas)                                                    NaN
avg(generation_fossil_brown_coal_lignite)                                     NaN
avg(generation_hydro_pumped_storage_aggregated)                               NaN
avg(generation_fossil_peat)                                                   NaN
avg(generation_wind_offshore)                                                 1.0
avg(generation_hydro_water_reservoir)                                         NaN
avg(total_load_actual)                                                        NaN
avg(generation_solar)                                                         NaN
avg(generation_geothermal)                                                    NaN
avg(generation_other)                                                         NaN
avg(price_actual)                                                             NaN
avg(generation_other_renewable)                                               NaN
avg(generation_nuclear)                                                       NaN
avg(generation_hydro_pumped_storage_consumption)                              NaN
avg(generation_biomass)                                                       NaN
avg(generation_marine)                                                        NaN
avg(generation_waste)                                                         NaN
avg(generation_fossil_hard_coal)                                              NaN
```

avg(generation_fossil_coal_derived_gas)	NaN
avg(generation_hydro_run_of_river_and_poundage)	NaN
avg(generation_fossil_oil_shale)	NaN
avg(generation_wind_onshore)	NaN
avg(generation_fossil_oil)	NaN

	avg(generation_hydro_water_reservoi
year	-0.1124
avg(generation_fossil_gas)	-0.7507
avg(generation_fossil_brown_coal_lignite)	-0.8404
avg(generation_hydro_pumped_storage_aggregated)	Na
avg(generation_fossil_peat)	Na
avg(generation_wind_offshore)	Na
avg(generation_hydro_water_reservoir)	1.0000
avg(total_load_actual)	-0.2433
avg(generation_solar)	-0.8571
avg(generation_geothermal)	Na
avg(generation_other)	0.2648
avg(price_actual)	-0.5063
avg(generation_other_renewable)	-0.2715
avg(generation_nuclear)	-0.1762
avg(generation_hydro_pumped_storage_consumption)	0.4884
avg(generation_biomass)	0.0432
avg(generation_marine)	Na
avg(generation_waste)	-0.3020
avg(generation_fossil_hard_coal)	-0.4429
avg(generation_fossil_coal_derived_gas)	Na
avg(generation_hydro_run_of_river_and_poundage)	0.7284
avg(generation_fossil_oil_shale)	Na
avg(generation_wind_onshore)	0.3456
avg(generation_fossil_oil)	-0.2646

	avg(total_load_actual) \
year	0.988030
avg(generation_fossil_gas)	0.772630
avg(generation_fossil_brown_coal_lignite)	-0.320710
avg(generation_hydro_pumped_storage_aggregated)	NaN
avg(generation_fossil_peat)	NaN
avg(generation_wind_offshore)	NaN
avg(generation_hydro_water_reservoir)	-0.243325
avg(total_load_actual)	1.000000
avg(generation_solar)	-0.266476
avg(generation_geothermal)	NaN
avg(generation_other)	-0.838801
avg(price_actual)	0.448746
avg(generation_other_renewable)	0.972060
avg(generation_nuclear)	-0.535851
avg(generation_hydro_pumped_storage_consumption)	-0.935592

avg(generation_biomass)	-0.807328
avg(generation_marine)	NaN
avg(generation_waste)	0.946351
avg(generation_fossil_hard_coal)	-0.586183
avg(generation_fossil_coal_derived_gas)	NaN
avg(generation_hydro_run_of_river_and_poundage)	0.435614
avg(generation_fossil_oil_shale)	NaN
avg(generation_wind_onshore)	0.513215
avg(generation_fossil_oil)	-0.701659

	avg(generation_solar) \
year	-0.376027
avg(generation_fossil_gas)	0.386652
avg(generation_fossil_brown_coal_lignite)	0.982195
avg(generation_hydro_pumped_storage_aggregated)	NaN
avg(generation_fossil_peat)	NaN
avg(generation_wind_offshore)	NaN
avg(generation_hydro_water_reservoir)	-0.857189
avg(total_load_actual)	-0.266476
avg(generation_solar)	1.000000
avg(generation_geothermal)	NaN
avg(generation_other)	0.077832
avg(price_actual)	0.143228
avg(generation_other_renewable)	-0.187746
avg(generation_nuclear)	0.567939
avg(generation_hydro_pumped_storage_consumption)	0.030115
avg(generation_biomass)	0.274732
avg(generation_marine)	NaN
avg(generation_waste)	-0.131752
avg(generation_fossil_hard_coal)	0.644033
avg(generation_fossil_coal_derived_gas)	NaN
avg(generation_hydro_run_of_river_and_poundage)	-0.976062
avg(generation_fossil_oil_shale)	NaN
avg(generation_wind_onshore)	-0.710615
avg(generation_fossil_oil)	0.525526

	avg(generation_geothermal) \
year	NaN
avg(generation_fossil_gas)	NaN
avg(generation_fossil_brown_coal_lignite)	NaN
avg(generation_hydro_pumped_storage_aggregated)	NaN
avg(generation_fossil_peat)	NaN
avg(generation_wind_offshore)	NaN
avg(generation_hydro_water_reservoir)	NaN
avg(total_load_actual)	NaN
avg(generation_solar)	NaN
avg(generation_geothermal)	1.0
avg(generation_other)	NaN

```
avg(price_actual)                                                    NaN
avg(generation_other_renewable)                                      NaN
avg(generation_nuclear)                                              NaN
avg(generation_hydro_pumped_storage_consumption)                     NaN
avg(generation_biomass)                                              NaN
avg(generation_marine)                                               NaN
avg(generation_waste)                                                NaN
avg(generation_fossil_hard_coal)                                     NaN
avg(generation_fossil_coal_derived_gas)                              NaN
avg(generation_hydro_run_of_river_and_poundage)                      NaN
avg(generation_fossil_oil_shale)                                     NaN
avg(generation_wind_onshore)                                         NaN
avg(generation_fossil_oil)                                           NaN

                                                    ...  \
year                                                ...
avg(generation_fossil_gas)                          ...
avg(generation_fossil_brown_coal_lignite)           ...
avg(generation_hydro_pumped_storage_aggregated)     ...
avg(generation_fossil_peat)                         ...
avg(generation_wind_offshore)                       ...
avg(generation_hydro_water_reservoir)               ...
avg(total_load_actual)                              ...
avg(generation_solar)                               ...
avg(generation_geothermal)                          ...
avg(generation_other)                               ...
avg(price_actual)                                   ...
avg(generation_other_renewable)                     ...
avg(generation_nuclear)                             ...
avg(generation_hydro_pumped_storage_consumption)    ...
avg(generation_biomass)                             ...
avg(generation_marine)                              ...
avg(generation_waste)                               ...
avg(generation_fossil_hard_coal)                    ...
avg(generation_fossil_coal_derived_gas)             ...
avg(generation_hydro_run_of_river_and_poundage)     ...
avg(generation_fossil_oil_shale)                    ...
avg(generation_wind_onshore)                        ...
avg(generation_fossil_oil)                          ...

                                                    avg(generation_hydro_pumped_storage_
year
avg(generation_fossil_gas)
avg(generation_fossil_brown_coal_lignite)
avg(generation_hydro_pumped_storage_aggregated)
avg(generation_fossil_peat)
avg(generation_wind_offshore)
avg(generation_hydro_water_reservoir)
```

```
avg(total_load_actual)
avg(generation_solar)
avg(generation_geothermal)
avg(generation_other)
avg(price_actual)
avg(generation_other_renewable)
avg(generation_nuclear)
avg(generation_hydro_pumped_storage_consumption)
avg(generation_biomass)
avg(generation_marine)
avg(generation_waste)
avg(generation_fossil_hard_coal)
avg(generation_fossil_coal_derived_gas)
avg(generation_hydro_run_of_river_and_poundage)
avg(generation_fossil_oil_shale)
avg(generation_wind_onshore)
avg(generation_fossil_oil)
```

	avg(generation_biomass)	\
year	-0.863918	
avg(generation_fossil_gas)	-0.678732	
avg(generation_fossil_brown_coal_lignite)	0.420684	
avg(generation_hydro_pumped_storage_aggregated)	NaN	
avg(generation_fossil_peat)	NaN	
avg(generation_wind_offshore)	NaN	
avg(generation_hydro_water_reservoir)	0.043260	
avg(total_load_actual)	-0.807328	
avg(generation_solar)	0.274732	
avg(generation_geothermal)	NaN	
avg(generation_other)	0.974787	
avg(price_actual)	0.154014	
avg(generation_other_renewable)	-0.910937	
avg(generation_nuclear)	0.037389	
avg(generation_hydro_pumped_storage_consumption)	0.582766	
avg(generation_biomass)	1.000000	
avg(generation_marine)	NaN	
avg(generation_waste)	-0.931177	
avg(generation_fossil_hard_coal)	0.875800	
avg(generation_fossil_coal_derived_gas)	NaN	
avg(generation_hydro_run_of_river_and_poundage)	-0.337062	
avg(generation_fossil_oil_shale)	NaN	
avg(generation_wind_onshore)	-0.088375	
avg(generation_fossil_oil)	0.951989	

	avg(generation_marine)	\
year	NaN	
avg(generation_fossil_gas)	NaN	
avg(generation_fossil_brown_coal_lignite)	NaN	

avg(generation_hydro_pumped_storage_aggregated)	NaN
avg(generation_fossil_peat)	NaN
avg(generation_wind_offshore)	NaN
avg(generation_hydro_water_reservoir)	NaN
avg(total_load_actual)	NaN
avg(generation_solar)	NaN
avg(generation_geothermal)	NaN
avg(generation_other)	NaN
avg(price_actual)	NaN
avg(generation_other_renewable)	NaN
avg(generation_nuclear)	NaN
avg(generation_hydro_pumped_storage_consumption)	NaN
avg(generation_biomass)	NaN
avg(generation_marine)	1.0
avg(generation_waste)	NaN
avg(generation_fossil_hard_coal)	NaN
avg(generation_fossil_coal_derived_gas)	NaN
avg(generation_hydro_run_of_river_and_poundage)	NaN
avg(generation_fossil_oil_shale)	NaN
avg(generation_wind_onshore)	NaN
avg(generation_fossil_oil)	NaN

	avg(generation_waste) \
year	0.950379
avg(generation_fossil_gas)	0.855207
avg(generation_fossil_brown_coal_lignite)	-0.240190
avg(generation_hydro_pumped_storage_aggregated)	NaN
avg(generation_fossil_peat)	NaN
avg(generation_wind_offshore)	NaN
avg(generation_hydro_water_reservoir)	-0.302046
avg(total_load_actual)	0.946351
avg(generation_solar)	-0.131752
avg(generation_geothermal)	NaN
avg(generation_other)	-0.968555
avg(price_actual)	0.205121
avg(generation_other_renewable)	0.995540
avg(generation_nuclear)	-0.237445
avg(generation_hydro_pumped_storage_consumption)	-0.833325
avg(generation_biomass)	-0.931177
avg(generation_marine)	NaN
avg(generation_waste)	1.000000
avg(generation_fossil_hard_coal)	-0.680025
avg(generation_fossil_coal_derived_gas)	NaN
avg(generation_hydro_run_of_river_and_poundage)	0.265321
avg(generation_fossil_oil_shale)	NaN
avg(generation_wind_onshore)	0.227072
avg(generation_fossil_oil)	-0.804868

	avg(generation_fossil_hard_coal)	\
year	-0.701993	
avg(generation_fossil_gas)	-0.241981	
avg(generation_fossil_brown_coal_lignite)	0.771253	
avg(generation_hydro_pumped_storage_aggregated)	NaN	
avg(generation_fossil_peat)	NaN	
avg(generation_wind_offshore)	NaN	
avg(generation_hydro_water_reservoir)	-0.442907	
avg(total_load_actual)	-0.586183	
avg(generation_solar)	0.644033	
avg(generation_geothermal)	NaN	
avg(generation_other)	0.746076	
avg(price_actual)	0.413356	
avg(generation_other_renewable)	-0.673309	
avg(generation_nuclear)	0.081910	
avg(generation_hydro_pumped_storage_consumption)	0.262458	
avg(generation_biomass)	0.875800	
avg(generation_marine)	NaN	
avg(generation_waste)	-0.680025	
avg(generation_fossil_hard_coal)	1.000000	
avg(generation_fossil_coal_derived_gas)	NaN	
avg(generation_hydro_run_of_river_and_poundage)	-0.631705	
avg(generation_fossil_oil_shale)	NaN	
avg(generation_wind_onshore)	-0.211109	
avg(generation_fossil_oil)	0.981258	

	avg(generation_fossil_coal_derived_g
year	
avg(generation_fossil_gas)	
avg(generation_fossil_brown_coal_lignite)	
avg(generation_hydro_pumped_storage_aggregated)	
avg(generation_fossil_peat)	
avg(generation_wind_offshore)	
avg(generation_hydro_water_reservoir)	
avg(total_load_actual)	
avg(generation_solar)	
avg(generation_geothermal)	
avg(generation_other)	
avg(price_actual)	
avg(generation_other_renewable)	
avg(generation_nuclear)	
avg(generation_hydro_pumped_storage_consumption)	
avg(generation_biomass)	
avg(generation_marine)	
avg(generation_waste)	
avg(generation_fossil_hard_coal)	
avg(generation_fossil_coal_derived_gas)	
avg(generation_hydro_run_of_river_and_poundage)	

```
avg(generation_fossil_oil_shale)
avg(generation_wind_onshore)
avg(generation_fossil_oil)

                                                    avg(generation_hydro_run_of_river_a
year
avg(generation_fossil_gas)
avg(generation_fossil_brown_coal_lignite)
avg(generation_hydro_pumped_storage_aggregated)
avg(generation_fossil_peat)
avg(generation_wind_offshore)
avg(generation_hydro_water_reservoir)
avg(total_load_actual)
avg(generation_solar)
avg(generation_geothermal)
avg(generation_other)
avg(price_actual)
avg(generation_other_renewable)
avg(generation_nuclear)
avg(generation_hydro_pumped_storage_consumption)
avg(generation_biomass)
avg(generation_marine)
avg(generation_waste)
avg(generation_fossil_hard_coal)
avg(generation_fossil_coal_derived_gas)
avg(generation_hydro_run_of_river_and_poundage)
avg(generation_fossil_oil_shale)
avg(generation_wind_onshore)
avg(generation_fossil_oil)

                                                    avg(generation_fossil_oil_shale)  \
year                                                                               NaN
avg(generation_fossil_gas)                                                         NaN
avg(generation_fossil_brown_coal_lignite)                                          NaN
avg(generation_hydro_pumped_storage_aggregated)                                    NaN
avg(generation_fossil_peat)                                                        NaN
avg(generation_wind_offshore)                                                      NaN
avg(generation_hydro_water_reservoir)                                              NaN
avg(total_load_actual)                                                             NaN
avg(generation_solar)                                                              NaN
avg(generation_geothermal)                                                         NaN
avg(generation_other)                                                              NaN
avg(price_actual)                                                                  NaN
avg(generation_other_renewable)                                                    NaN
avg(generation_nuclear)                                                            NaN
avg(generation_hydro_pumped_storage_consumption)                                   NaN
avg(generation_biomass)                                                            NaN
avg(generation_marine)                                                             NaN
```

avg(generation_waste)	NaN
avg(generation_fossil_hard_coal)	NaN
avg(generation_fossil_coal_derived_gas)	NaN
avg(generation_hydro_run_of_river_and_poundage)	NaN
avg(generation_fossil_oil_shale)	1.0
avg(generation_wind_onshore)	NaN
avg(generation_fossil_oil)	NaN

	avg(generation_wind_onshore) \
year	0.514543
avg(generation_fossil_gas)	-0.067438
avg(generation_fossil_brown_coal_lignite)	-0.608724
avg(generation_hydro_pumped_storage_aggregated)	NaN
avg(generation_fossil_peat)	NaN
avg(generation_wind_offshore)	NaN
avg(generation_hydro_water_reservoir)	0.345629
avg(total_load_actual)	0.513215
avg(generation_solar)	-0.710615
avg(generation_geothermal)	NaN
avg(generation_other)	-0.017169
avg(price_actual)	0.594049
avg(generation_other_renewable)	0.317127
avg(generation_nuclear)	-0.982664
avg(generation_hydro_pumped_storage_consumption)	-0.511716
avg(generation_biomass)	-0.088375
avg(generation_marine)	NaN
avg(generation_waste)	0.227072
avg(generation_fossil_hard_coal)	-0.211109
avg(generation_fossil_coal_derived_gas)	NaN
avg(generation_hydro_run_of_river_and_poundage)	0.828306
avg(generation_fossil_oil_shale)	NaN
avg(generation_wind_onshore)	1.000000
avg(generation_fossil_oil)	-0.186019

	avg(generation_fossil_oil)
year	-0.796692
avg(generation_fossil_gas)	-0.424401
avg(generation_fossil_brown_coal_lignite)	0.661934
avg(generation_hydro_pumped_storage_aggregated)	NaN
avg(generation_fossil_peat)	NaN
avg(generation_wind_offshore)	NaN
avg(generation_hydro_water_reservoir)	-0.264612
avg(total_load_actual)	-0.701659
avg(generation_solar)	0.525526
avg(generation_geothermal)	NaN
avg(generation_other)	0.859697
avg(price_actual)	0.308487
avg(generation_other_renewable)	-0.794180

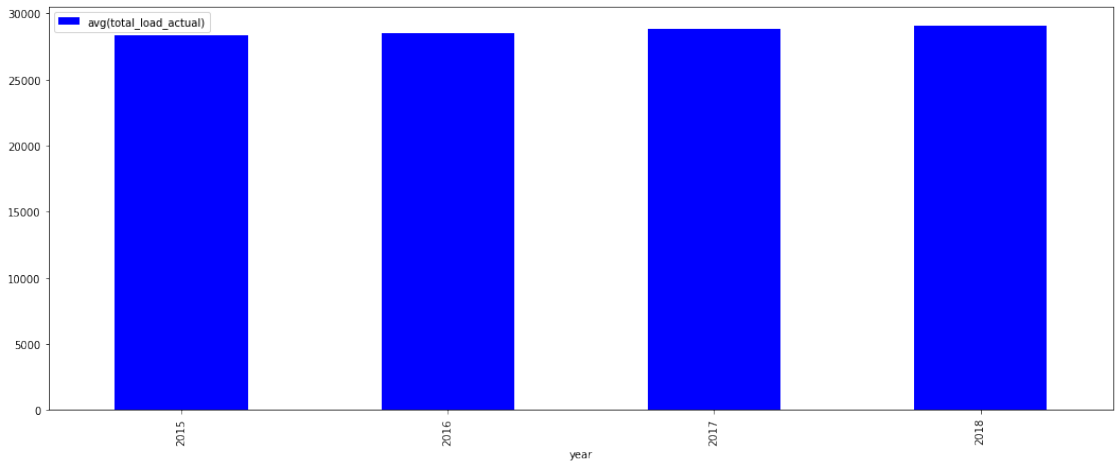
```
avg(generation_nuclear)                                    0.084624
avg(generation_hydro_pumped_storage_consumption)           0.409152
avg(generation_biomass)                                    0.951989
avg(generation_marine)                                          NaN
avg(generation_waste)                                     -0.804868
avg(generation_fossil_hard_coal)                           0.981258
avg(generation_fossil_coal_derived_gas)                         NaN
avg(generation_hydro_run_of_river_and_poundage)           -0.545330
avg(generation_fossil_oil_shale)                                NaN
avg(generation_wind_onshore)                              -0.186019
avg(generation_fossil_oil)                                 1.000000

[24 rows x 24 columns]
```

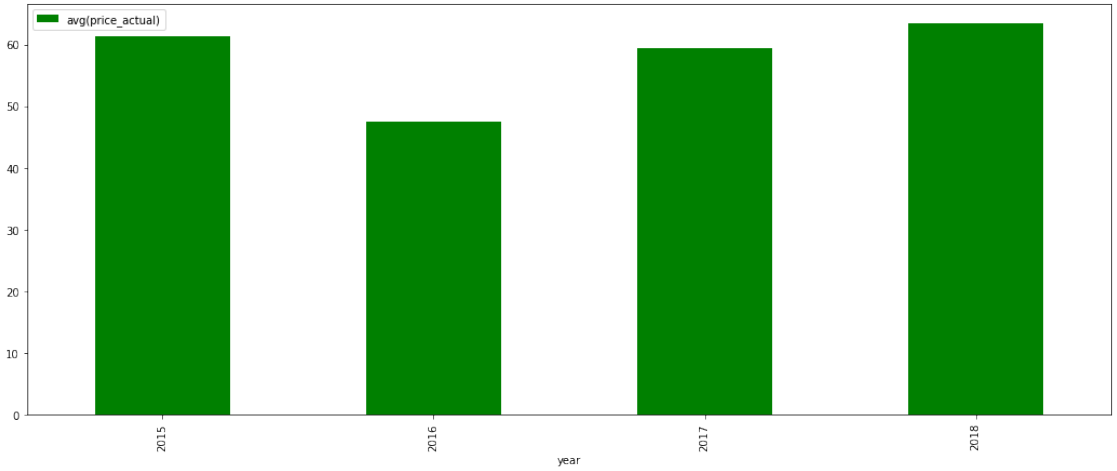
The following columns can be ignored: - generation_hydro_pumped_storage_aggregated - generation_fossil_peat - generation_wind_offshore - generation_geothermal - generation_marine - generation_fossil_coal_derived_gas - generation_fossil_oil_shale

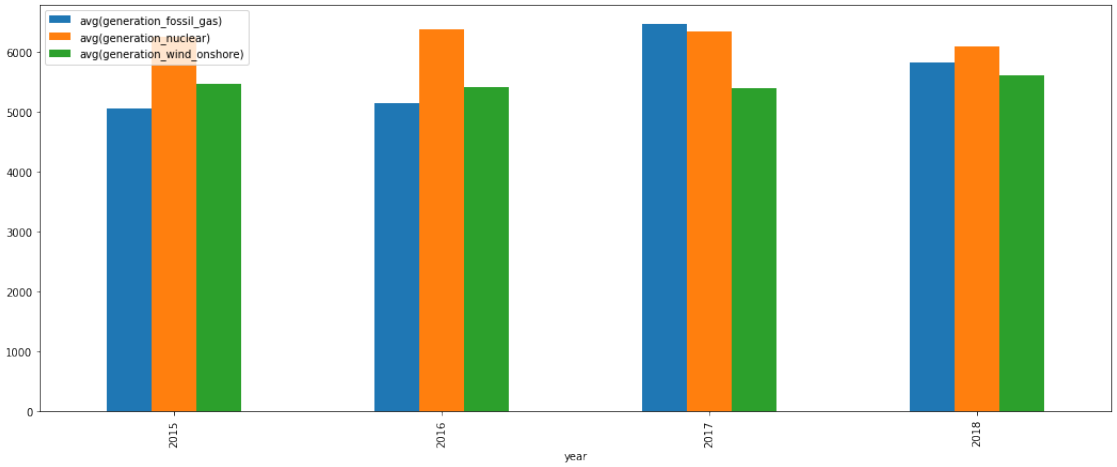
Using the correlations listed above, strong ones will be visualized and analyzed

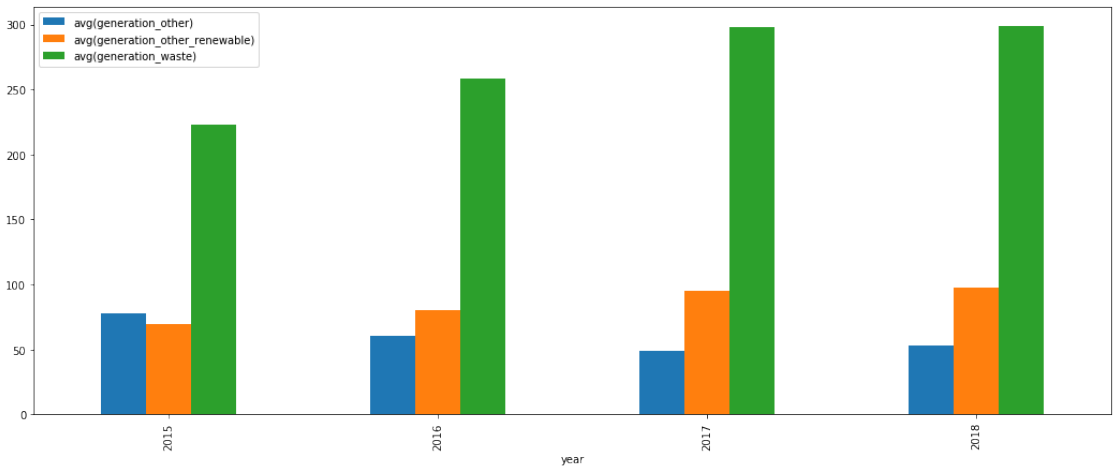
```
In [62]: graph_df = test_agg.select('*').toPandas()
         plt.rcParams["figure.figsize"] = 18,7
         graph_df.plot.bar(x='year', y='avg(total_load_actual)', color='blue')
         graph_df.plot.bar(x='year', y='avg(price_actual)', color='green')
         graph_df.plot.bar(x='year',
                           y=['avg(generation_fossil_gas)',
                              'avg(generation_nuclear)',
                              'avg(generation_wind_onshore)'],
                           sharex=True)
         graph_df.plot.bar(x='year',
                           y=['avg(generation_other)',
                              'avg(generation_other_renewable)',
                              'avg(generation_waste)'],
                           sharex=True)
```

```
Out[62]: <matplotlib.axes._subplots.AxesSubplot at 0x1834db10>
```




avg(price_actual)
60
50
40
30
20
10
0
2015
2016
2017
2018
year


avg(generation_fossil_gas)
avg(generation_nuclear)
avg(generation_wind_onshore)
6000
5000
4000
3000
2000
1000
0
2015
2016
2017
2018
year


avg(generation_other)
avg(generation_other_renewable)
avg(generation_waste)
300
250
200
150
100
50
0
2015
2016
2017
2018
year

The total load varies very little by year, meaning that overall the system is undertaking approximately the same amount of work each year. This is important to note beacuse any patterns of increase or decrease we find within a Generation column, means that it is generally cancelled out by the decrease or increase in another Generation column

There was a substantial price difference in 2016, being much lower than the surrounding years.

Looking at a subset of Generation columns we can start to see how they each begin to follow their own pattern, and some of them draw much much more power than others.

The following generate an incredibly high MegaWatt drain: - Generation Fossil Gas - Generation Nuclear - Generation Wind Onshore

The following have a very small draw in comparison: - Generation Other - Generation Other Renewable - Generation Waste

4.5 Generation_ By Month

Looking at the same Generation columns as we did above, but this time at a Month granularity as opposed to year. We are also ignoring any column that was deemed unimportant

```
In [63]: test_agg = joined_df.groupBy(f.month('time'))\
             .agg({"price_actual" : "avg", \
                   "total_load_actual" : "avg",\
                   "generation_biomass" : "avg",\
                   "generation_fossil_brown_coal_lignite" : "avg",\
                   "generation_fossil_gas" : "avg",\
                   "generation_fossil_hard_coal" : "avg",\
                   "generation_fossil_oil" : "avg",\
                   "generation_hydro_pumped_storage_consumption" : "avg",\
                   "generation_hydro_run_of_river_and_poundage" : "avg",\
                   "generation_hydro_water_reservoir" : "avg",\
                   "generation_nuclear" : "avg",\
                   "generation_other" : "avg",\
                   "generation_other_renewable" : "avg",\
                   "generation_solar" : "avg",\
                   "generation_waste" : "avg",\
                   "generation_wind_onshore" : "avg",\
                  })\
             .withColumnRenamed('month(time)', 'month')\
             .orderBy('month')

In [64]: correlation_matrix(test_agg, test_agg.columns)

Out[64]:                                                    month  \
         month                                           1.000000
         avg(generation_fossil_gas)                      0.721932
         avg(generation_fossil_brown_coal_lignite)       0.627535
         avg(generation_hydro_water_reservoir)          -0.801728
         avg(total_load_actual)                         -0.250626
```

```
avg(generation_solar)                              -0.152388
avg(generation_other)                               0.036723
avg(price_actual)                                   0.688920
avg(generation_other_renewable)                     0.502845
avg(generation_nuclear)                            -0.250136
avg(generation_hydro_pumped_storage_consumption)   -0.496155
avg(generation_biomass)                            -0.007149
avg(generation_waste)                               0.534335
avg(generation_fossil_hard_coal)                    0.566349
avg(generation_hydro_run_of_river_and_poundage)    -0.707855
avg(generation_wind_onshore)                       -0.610899
avg(generation_fossil_oil)                         -0.009340

                                                   avg(generation_fossil_gas)  \
month                                                                0.721932
avg(generation_fossil_gas)                                           1.000000
avg(generation_fossil_brown_coal_lignite)                            0.893011
avg(generation_hydro_water_reservoir)                               -0.815705
avg(total_load_actual)                                               0.326877
avg(generation_solar)                                               -0.127170
avg(generation_other)                                                0.204300
avg(price_actual)                                                    0.864872
avg(generation_other_renewable)                                      0.400812
avg(generation_nuclear)                                             -0.266553
avg(generation_hydro_pumped_storage_consumption)                    -0.462019
avg(generation_biomass)                                              0.408114
avg(generation_waste)                                                0.713947
avg(generation_fossil_hard_coal)                                     0.915238
avg(generation_hydro_run_of_river_and_poundage)                     -0.766604
avg(generation_wind_onshore)                                        -0.501046
avg(generation_fossil_oil)                                           0.339675

                                                   avg(generation_fossil_brown_coal_lig
month                                                                              0.6
avg(generation_fossil_gas)                                                         0.8
avg(generation_fossil_brown_coal_lignite)                                          1.0
avg(generation_hydro_water_reservoir)                                             -0.8
avg(total_load_actual)                                                             0.3
avg(generation_solar)                                                             -0.1
avg(generation_other)                                                              0.1
avg(price_actual)                                                                  0.9
avg(generation_other_renewable)                                                    0.4
avg(generation_nuclear)                                                           -0.0
avg(generation_hydro_pumped_storage_consumption)                                  -0.4
avg(generation_biomass)                                                            0.5
avg(generation_waste)                                                              0.8
avg(generation_fossil_hard_coal)                                                   0.9
avg(generation_hydro_run_of_river_and_poundage)                                   -0.8
```

avg(generation_wind_onshore)	-0.5
avg(generation_fossil_oil)	0.4

	avg(generation_hydro_water_reservoi
month	-0.8017
avg(generation_fossil_gas)	-0.8157
avg(generation_fossil_brown_coal_lignite)	-0.8594
avg(generation_hydro_water_reservoir)	1.0000
avg(total_load_actual)	-0.0444
avg(generation_solar)	-0.0478
avg(generation_other)	-0.2053
avg(price_actual)	-0.8681
avg(generation_other_renewable)	-0.3888
avg(generation_nuclear)	-0.0000
avg(generation_hydro_pumped_storage_consumption)	0.6504
avg(generation_biomass)	-0.4309
avg(generation_waste)	-0.8508
avg(generation_fossil_hard_coal)	-0.8355
avg(generation_hydro_run_of_river_and_poundage)	0.9600
avg(generation_wind_onshore)	0.7518
avg(generation_fossil_oil)	-0.3491

	avg(total_load_actual) \
month	-0.250626
avg(generation_fossil_gas)	0.326877
avg(generation_fossil_brown_coal_lignite)	0.351370
avg(generation_hydro_water_reservoir)	-0.044412
avg(total_load_actual)	1.000000
avg(generation_solar)	0.014620
avg(generation_other)	0.388074
avg(price_actual)	0.295006
avg(generation_other_renewable)	-0.327580
avg(generation_nuclear)	0.281043
avg(generation_hydro_pumped_storage_consumption)	0.061434
avg(generation_biomass)	0.664148
avg(generation_waste)	0.289135
avg(generation_fossil_hard_coal)	0.464990
avg(generation_hydro_run_of_river_and_poundage)	-0.062771
avg(generation_wind_onshore)	0.218531
avg(generation_fossil_oil)	0.626397

	avg(generation_solar) \
month	-0.152388
avg(generation_fossil_gas)	-0.127170
avg(generation_fossil_brown_coal_lignite)	-0.162395
avg(generation_hydro_water_reservoir)	-0.047851
avg(total_load_actual)	0.014620
avg(generation_solar)	1.000000

avg(generation_other)	-0.299850
avg(price_actual)	-0.304441
avg(generation_other_renewable)	-0.474513
avg(generation_nuclear)	-0.095104
avg(generation_hydro_pumped_storage_consumption)	-0.725989
avg(generation_biomass)	0.067916
avg(generation_waste)	0.155315
avg(generation_fossil_hard_coal)	-0.128250
avg(generation_hydro_run_of_river_and_poundage)	0.047322
avg(generation_wind_onshore)	-0.554381
avg(generation_fossil_oil)	0.294699

	avg(generation_other) \
month	0.036723
avg(generation_fossil_gas)	0.204300
avg(generation_fossil_brown_coal_lignite)	0.223083
avg(generation_hydro_water_reservoir)	-0.205382
avg(total_load_actual)	0.388074
avg(generation_solar)	-0.299850
avg(generation_other)	1.000000
avg(price_actual)	0.344915
avg(generation_other_renewable)	-0.048247
avg(generation_nuclear)	0.300006
avg(generation_hydro_pumped_storage_consumption)	0.133208
avg(generation_biomass)	0.702590
avg(generation_waste)	0.110818
avg(generation_fossil_hard_coal)	0.301946
avg(generation_hydro_run_of_river_and_poundage)	-0.297569
avg(generation_wind_onshore)	0.314799
avg(generation_fossil_oil)	0.591783

	avg(price_actual) \
month	0.688920
avg(generation_fossil_gas)	0.864872
avg(generation_fossil_brown_coal_lignite)	0.935969
avg(generation_hydro_water_reservoir)	-0.868178
avg(total_load_actual)	0.295006
avg(generation_solar)	-0.304441
avg(generation_other)	0.344915
avg(price_actual)	1.000000
avg(generation_other_renewable)	0.568724
avg(generation_nuclear)	0.009509
avg(generation_hydro_pumped_storage_consumption)	-0.294214
avg(generation_biomass)	0.494994
avg(generation_waste)	0.748662
avg(generation_fossil_hard_coal)	0.928118
avg(generation_hydro_run_of_river_and_poundage)	-0.850738
avg(generation_wind_onshore)	-0.419002

```
avg(generation_fossil_oil)                                    0.324440

                                                  avg(generation_other_renewable)  \
month                                                                    0.502845
avg(generation_fossil_gas)                                               0.400812
avg(generation_fossil_brown_coal_lignite)                                0.490832
avg(generation_hydro_water_reservoir)                                   -0.388870
avg(total_load_actual)                                                  -0.327580
avg(generation_solar)                                                   -0.474513
avg(generation_other)                                                   -0.048247
avg(price_actual)                                                        0.568724
avg(generation_other_renewable)                                          1.000000
avg(generation_nuclear)                                                 -0.445214
avg(generation_hydro_pumped_storage_consumption)                         0.027449
avg(generation_biomass)                                                 -0.104189
avg(generation_waste)                                                    0.094508
avg(generation_fossil_hard_coal)                                         0.403654
avg(generation_hydro_run_of_river_and_poundage)                         -0.360714
avg(generation_wind_onshore)                                            -0.196771
avg(generation_fossil_oil)                                              -0.226934

                                                  avg(generation_nuclear)  \
month                                                           -0.250136
avg(generation_fossil_gas)                                      -0.266553
avg(generation_fossil_brown_coal_lignite)                       -0.035546
avg(generation_hydro_water_reservoir)                           -0.000075
avg(total_load_actual)                                           0.281043
avg(generation_solar)                                           -0.095104
avg(generation_other)                                            0.300006
avg(price_actual)                                                0.009509
avg(generation_other_renewable)                                 -0.445214
avg(generation_nuclear)                                          1.000000
avg(generation_hydro_pumped_storage_consumption)                 0.213965
avg(generation_biomass)                                          0.236782
avg(generation_waste)                                            0.303606
avg(generation_fossil_hard_coal)                                -0.058392
avg(generation_hydro_run_of_river_and_poundage)                 -0.150658
avg(generation_wind_onshore)                                     0.187460
avg(generation_fossil_oil)                                       0.268101

                                                  avg(generation_hydro_pumped_storage_
month
avg(generation_fossil_gas)
avg(generation_fossil_brown_coal_lignite)
avg(generation_hydro_water_reservoir)
avg(total_load_actual)
avg(generation_solar)
avg(generation_other)
```

```
avg(price_actual)
avg(generation_other_renewable)
avg(generation_nuclear)
avg(generation_hydro_pumped_storage_consumption)
avg(generation_biomass)
avg(generation_waste)
avg(generation_fossil_hard_coal)
avg(generation_hydro_run_of_river_and_poundage)
avg(generation_wind_onshore)
avg(generation_fossil_oil)
```

	avg(generation_biomass)	\
month	-0.007149	
avg(generation_fossil_gas)	0.408114	
avg(generation_fossil_brown_coal_lignite)	0.545007	
avg(generation_hydro_water_reservoir)	-0.430973	
avg(total_load_actual)	0.664148	
avg(generation_solar)	0.067916	
avg(generation_other)	0.702590	
avg(price_actual)	0.494994	
avg(generation_other_renewable)	-0.104189	
avg(generation_nuclear)	0.236782	
avg(generation_hydro_pumped_storage_consumption)	-0.219000	
avg(generation_biomass)	1.000000	
avg(generation_waste)	0.396271	
avg(generation_fossil_hard_coal)	0.619570	
avg(generation_hydro_run_of_river_and_poundage)	-0.490738	
avg(generation_wind_onshore)	-0.071424	
avg(generation_fossil_oil)	0.850527	

	avg(generation_waste)	\
month	0.534335	
avg(generation_fossil_gas)	0.713947	
avg(generation_fossil_brown_coal_lignite)	0.804039	
avg(generation_hydro_water_reservoir)	-0.850827	
avg(total_load_actual)	0.289135	
avg(generation_solar)	0.155315	
avg(generation_other)	0.110818	
avg(price_actual)	0.748662	
avg(generation_other_renewable)	0.094508	
avg(generation_nuclear)	0.303606	
avg(generation_hydro_pumped_storage_consumption)	-0.611523	
avg(generation_biomass)	0.396271	
avg(generation_waste)	1.000000	
avg(generation_fossil_hard_coal)	0.795894	
avg(generation_hydro_run_of_river_and_poundage)	-0.869397	
avg(generation_wind_onshore)	-0.696604	
avg(generation_fossil_oil)	0.377811	

	avg(generation_fossil_hard_coal)	\
month	0.566349	
avg(generation_fossil_gas)	0.915238	
avg(generation_fossil_brown_coal_lignite)	0.976464	
avg(generation_hydro_water_reservoir)	-0.835560	
avg(total_load_actual)	0.464990	
avg(generation_solar)	-0.128250	
avg(generation_other)	0.301946	
avg(price_actual)	0.928118	
avg(generation_other_renewable)	0.403654	
avg(generation_nuclear)	-0.058392	
avg(generation_hydro_pumped_storage_consumption)	-0.426933	
avg(generation_biomass)	0.619570	
avg(generation_waste)	0.795894	
avg(generation_fossil_hard_coal)	1.000000	
avg(generation_hydro_run_of_river_and_poundage)	-0.827654	
avg(generation_wind_onshore)	-0.486231	
avg(generation_fossil_oil)	0.471489	

	avg(generation_hydro_run_of_river_a
month	
avg(generation_fossil_gas)	
avg(generation_fossil_brown_coal_lignite)	
avg(generation_hydro_water_reservoir)	
avg(total_load_actual)	
avg(generation_solar)	
avg(generation_other)	
avg(price_actual)	
avg(generation_other_renewable)	
avg(generation_nuclear)	
avg(generation_hydro_pumped_storage_consumption)	
avg(generation_biomass)	
avg(generation_waste)	
avg(generation_fossil_hard_coal)	
avg(generation_hydro_run_of_river_and_poundage)	
avg(generation_wind_onshore)	
avg(generation_fossil_oil)	

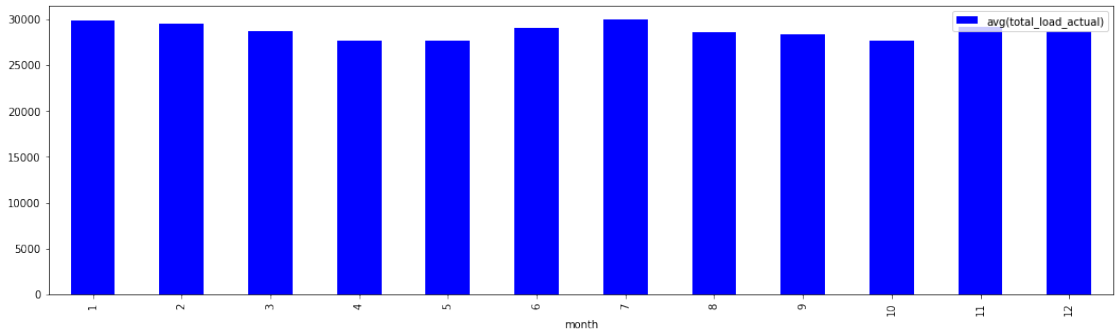
	avg(generation_wind_onshore)	\
month	-0.610899	
avg(generation_fossil_gas)	-0.501046	
avg(generation_fossil_brown_coal_lignite)	-0.542033	
avg(generation_hydro_water_reservoir)	0.751825	
avg(total_load_actual)	0.218531	
avg(generation_solar)	-0.554381	
avg(generation_other)	0.314799	
avg(price_actual)	-0.419002	

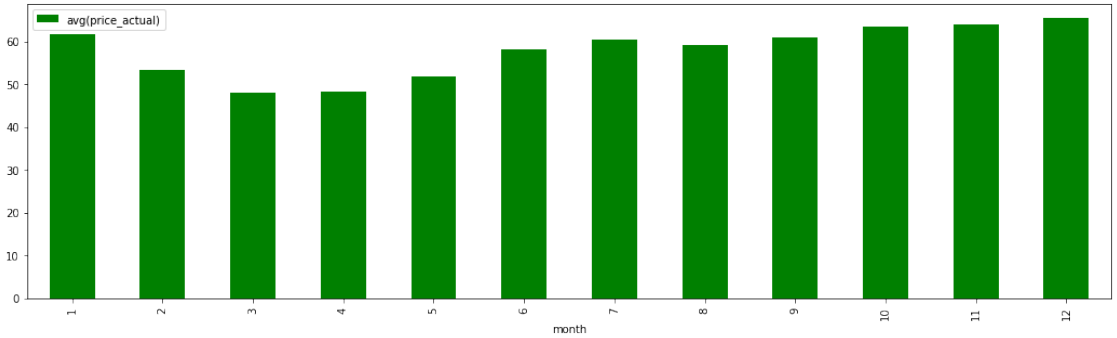
avg(generation_other_renewable)	-0.196771
avg(generation_nuclear)	0.187460
avg(generation_hydro_pumped_storage_consumption)	0.925689
avg(generation_biomass)	-0.071424
avg(generation_waste)	-0.696604
avg(generation_fossil_hard_coal)	-0.486231
avg(generation_hydro_run_of_river_and_poundage)	0.670355
avg(generation_wind_onshore)	1.000000
avg(generation_fossil_oil)	-0.159446

	avg(generation_fossil_oil)
month	-0.009340
avg(generation_fossil_gas)	0.339675
avg(generation_fossil_brown_coal_lignite)	0.429313
avg(generation_hydro_water_reservoir)	-0.349102
avg(total_load_actual)	0.626397
avg(generation_solar)	0.294699
avg(generation_other)	0.591783
avg(price_actual)	0.324440
avg(generation_other_renewable)	-0.226934
avg(generation_nuclear)	0.268101
avg(generation_hydro_pumped_storage_consumption)	-0.376654
avg(generation_biomass)	0.850527
avg(generation_waste)	0.377811
avg(generation_fossil_hard_coal)	0.471489
avg(generation_hydro_run_of_river_and_poundage)	-0.409427
avg(generation_wind_onshore)	-0.159446
avg(generation_fossil_oil)	1.000000

In [66]:
```
graph_df = test_agg.select('*').toPandas()
plt.rcParams["figure.figsize"] = 18,5
graph_df.plot.bar(x='month', y='avg(total_load_actual)', color='blue')
graph_df.plot.bar(x='month', y='avg(price_actual)', color='green')
```

Out[66]: <matplotlib.axes._subplots.AxesSubplot at 0x18648d10>



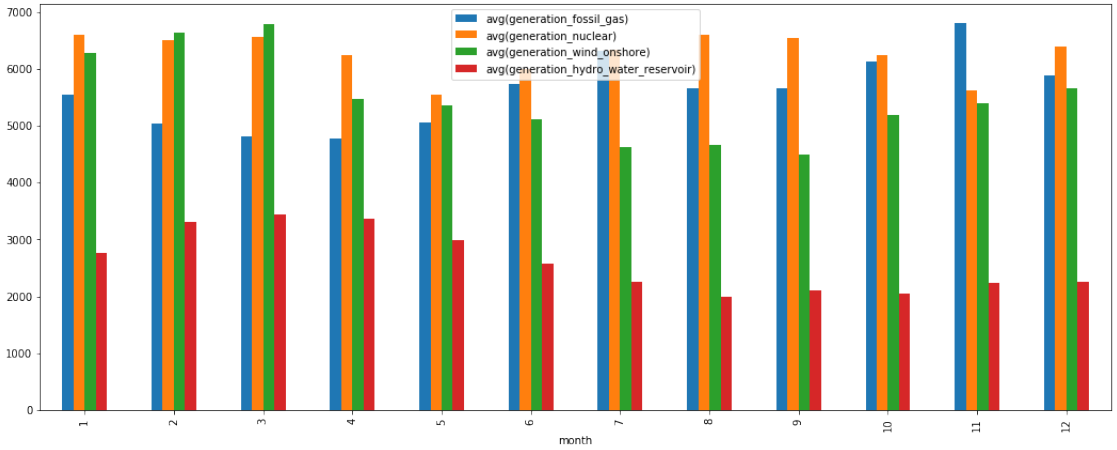


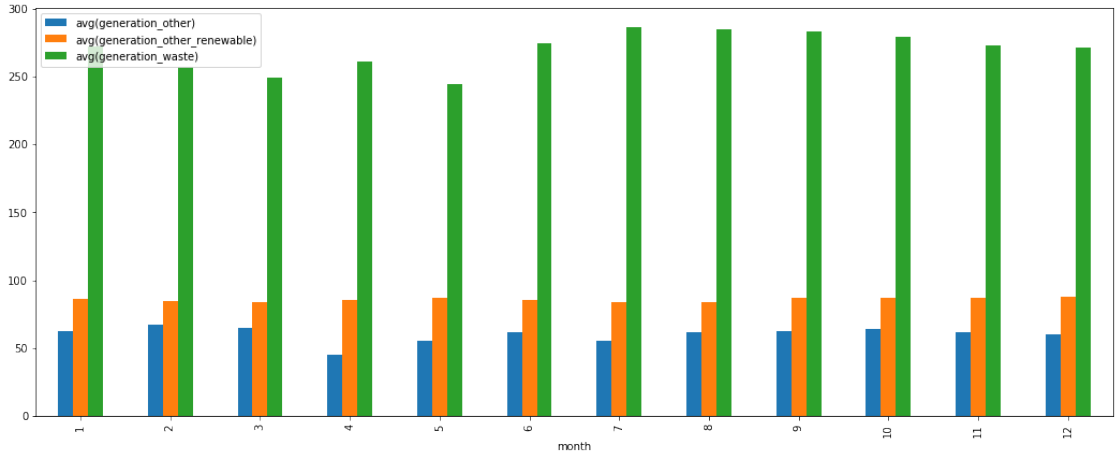
Looking by Month we can see a stronger indication of pattern, especially when looking by price. March and April being the cheapest month for the customers; December and January being the most expensive.

The total load remains generally consistent but the price and load do not seem to strongly influence one another.

```
In [68]: plt.rcParams["figure.figsize"] = 18,7

         graph_df.plot.bar(x='month',
                           y=['avg(generation_fossil_gas)',
                              'avg(generation_nuclear)',
                              'avg(generation_wind_onshore)',
                              'avg(generation_hydro_water_reservoir)'],
                           sharex=True)
         graph_df.plot.bar(x='month',
                           y=['avg(generation_other)',
                              'avg(generation_other_renewable)',
                              'avg(generation_waste)'],
                           sharex=True)
```

```
Out[68]: <matplotlib.axes._subplots.AxesSubplot at 0x1561f110>
```

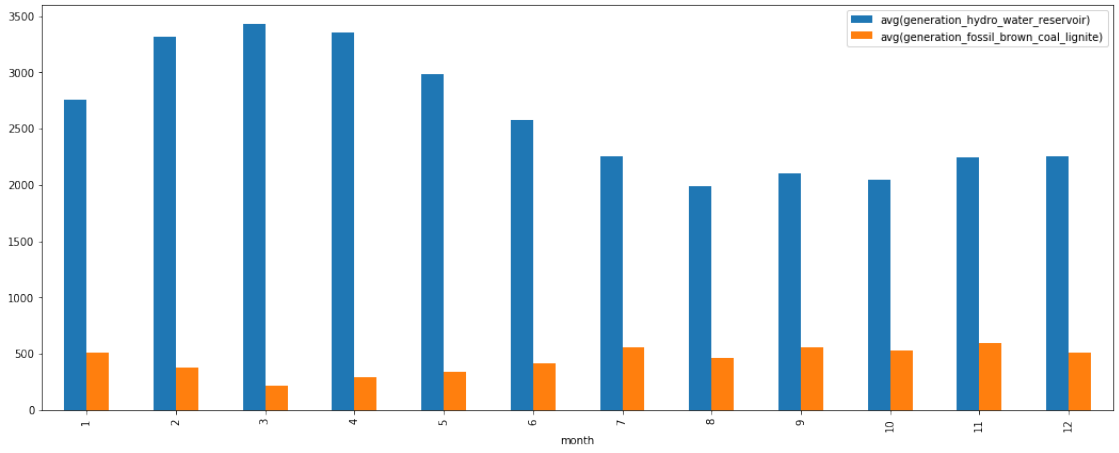




Increasing our time granularity to Month we can see stronger patterns begin to emerge. It's interesting to note that some Generation columns have a strong correlation to one another as well.

```
In [70]: graph_df.plot.bar(x='month',
                           y=['avg(generation_hydro_water_reservoir)',
                              'avg(generation_fossil_brown_coal_lignite)'],
                           sharex=True)
```

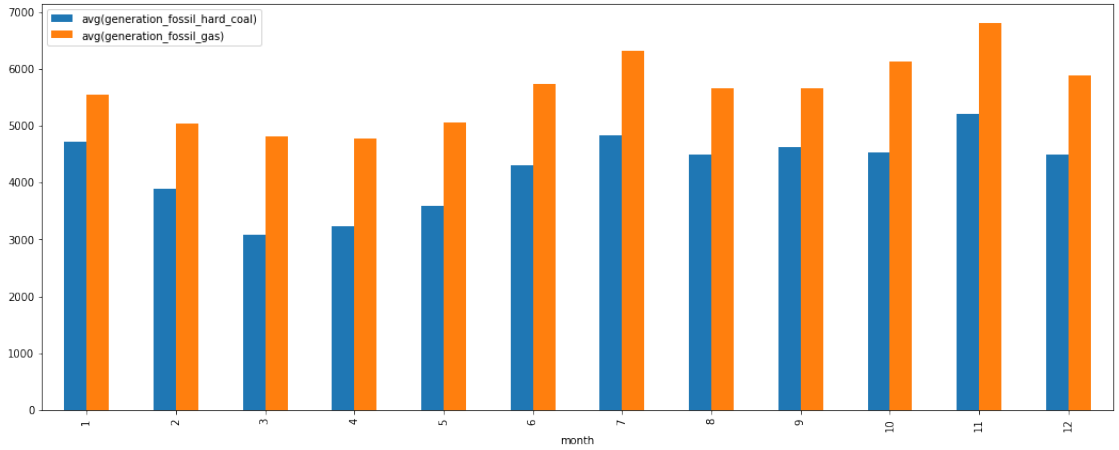
```
Out[70]: <matplotlib.axes._subplots.AxesSubplot at 0x16b695d0>
```



There is a very strong negative correlation between Hydro Water Reservoir and Fossil Brown Coal Lignite. I don't have enough domain knowledge about this subject to know whether or not this is to be expected

```
In [71]: graph_df.plot.bar(x='month',
                           y=['avg(generation_fossil_hard_coal)',
                              'avg(generation_fossil_gas)'],
                           sharex=True)
```

```
Out[71]: <matplotlib.axes._subplots.AxesSubplot at 0x16b5a4f0>
```



Others Generation columns have a very strong Generation columns that we can expect, such as two separate Fossil Generation columns

5 Does hour of day affect Energy Price or Load?

Going to an even deeper level of time granularity, how does the hour of day affect energy price or load?

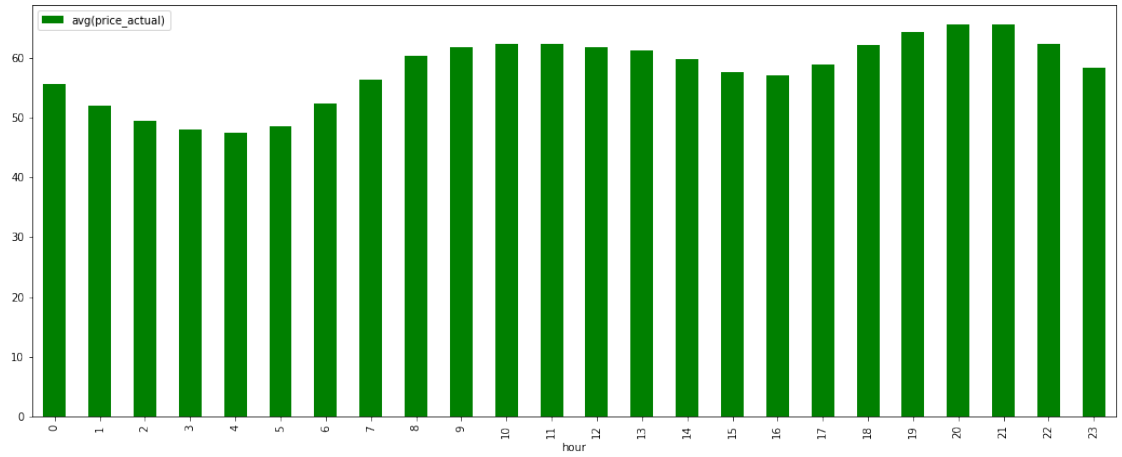
```
In [17]: test_agg = joined_df.groupBy(f.hour('time'))\
             .agg({"total_load_actual" : "avg", \
                   "price_actual" : "avg"})\
             .withColumnRenamed('hour(time)', 'hour')\
             .orderBy('hour(time)')
         test_agg.show(10,False)
```

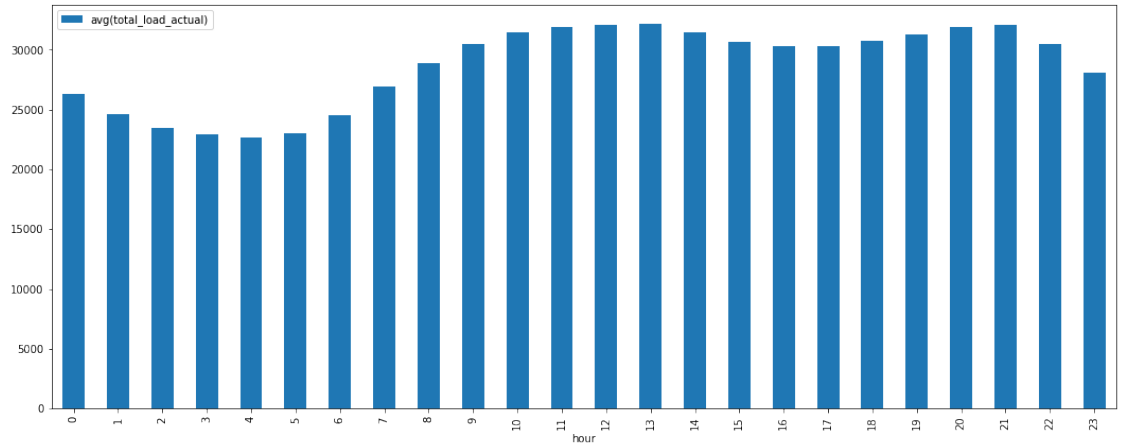
```
+----+------------------+---------------------+
|hour|avg(price_actual) |avg(total_load_actual)|
+----+------------------+---------------------+
|0   |55.54451529921473 |26289.17059301381    |
|1   |51.87812483076095 |24592.89371784457    |
|2   |49.380810010764236|23437.54036598493    |
|3   |47.999081124714095|22921.079752121783   |
|4   |47.443314939434686|22708.48923283984    |
|5   |48.44930890475544 |23000.831356160692   |
|6   |52.360210908113935|24573.628559914025   |
```

```
|7   |56.37116581701105 |26974.129768941428    |
|8   |60.264768123667416|28852.69021231139     |
|9   |61.81943133616112 |30442.86859954781     |
+----+------------------+----------------------+
only showing top 10 rows
```

```
In [41]: graph_df = test_agg.select('*').toPandas()
         plt.rcParams["figure.figsize"] = 18,7
         graph_df.plot.bar(x='hour', y='avg(price_actual)', color='green')
         graph_df.plot.bar(x='hour', y='avg(total_load_actual)')
```

```
Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x1e158410>
```






```
In [18]: correlation_matrix(test_agg, test_agg.columns)
```

Out[18]:

	hour	avg(price_actual)	avg(total_load_actual)
hour	1.000000	0.743679	0.741345
avg(price_actual)	0.743679	1.000000	0.948137
avg(total_load_actual)	0.741345	0.948137	1.000000

Total load on the system and actual price very strongly with the hour of day. Peak hour being the afternoon and late night, and lowest hours being 3-5 am

6 Conclusions

Looking through the data we've found a number of strong correlations within the data

We only had a brief introduction into machine learning within our CS 490, but time-series analysis would be a strong contender for this dataset as it would allow us to predict upcoming prices, load usage and energy generation for upcoming years and months