

Dutch Energy (2010 – 2019)

Definition

Project Overview:

Energy is in the core of our survival and development as human being. Understanding how a nation consume its energy and what are the trends of such consumption is vital for future planning for utility providers.

This case study will look at the electrical and gas consumption in the Netherlands for the past 9 years, from 2010 to 2019. Enexis, Liander, and Stedin are the three major network administrators of the Netherlands and, together, they provide energy to nearly the entire country.



Year	Population	Yearly % Change
2019	17,097,130	0.22 %
2018	17,059,560	0.22 %
2017	17,021,347	0.24 %
2016	16,981,295	0.25 %
2015	16,938,499	0.30 %
2010	16,682,917	0.38 %

Map of the Netherlands & Population Trend from 2010 to 2019

Problem Statement:

- In this project I will be addressing the following business questions:
- How are the smart meters spreading?
- What is the trend of solar panel installation , ie homes -produced energy
- Other insights into trends for different network administrators
- What is the energy consumption the next year?

Analysis

Loading Data:

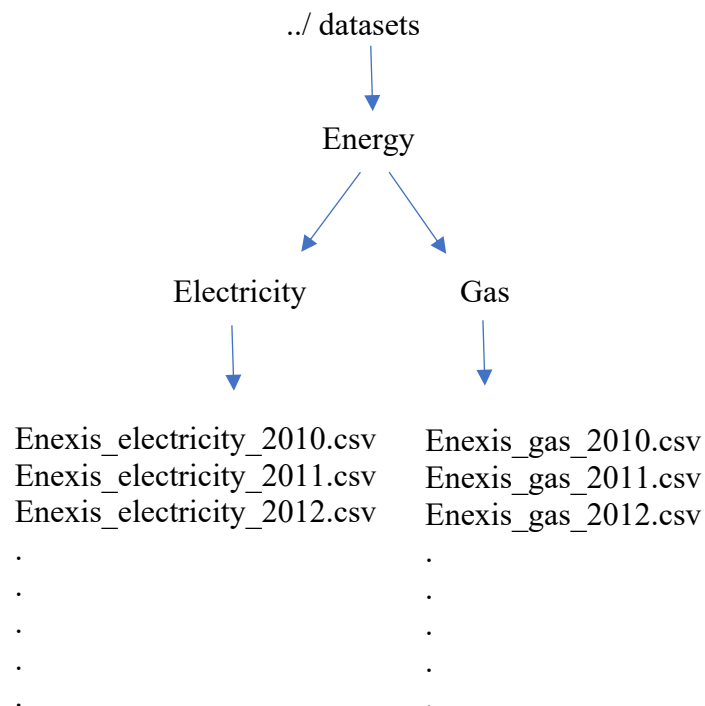
The data is obtained from Kaggle on the following [link](#). The file is in a zip format named `dutch_energy.zip`.

As first step , following library import, I have created list of functions that unzip the file , explore the content and rename each extracted file based on the below naming structure:

Network_administrator + Energy_type + year

For example, file containing electricity consumption for the year 2013 belonging to Enexis will be called: `Enexis_electricity_2013`

The functions also create folders, path to save all extracted csv. The final folder structure of extracted data is as follow:



Data Exploration:

On average extracted csv file for electricity and gas have number of rows(observations) between 80,000 to 120,000 representing the various connection for particular zipcodes under the network administrator for the specific year. Every entry describes at least 10 connections.

There are 14 features (columns) in each file as follow:

1- net_manager	Regional network administrator (Enexis, Liander, Stedin)
2- purchase_area	Code of the area where the energy is purchased
3- street	Street name
4- zipcode_from	Zipcode of range covered – start (4 numbers and 2 letters)
5- zipcode_to	Zipcode of range covered – finish 4 numbers and 2 letters)
6- city	City Name
7- num_connections	Number of connections in the range of zipcodes
8- delivery_perc	Percentage of electricity or gas consumed. The number will decrease if consumer give back energy to the grid - for example if solar panels have been installed.
9- perc_of_active_connections	Percentage of active connections in the zipcode range
10- type_of_connection	Main type of connection in the zipcode range. For electricity it is No of fuses X amount of Amps. For gas it is G4, G6, G10, G16, G25
11- type_conn_perc	Percentage of presence of the main type of connection in the zipcode range
12- annual_consume	Annual consumption. Electricity (kwh), Gas(m3)
13- annual_consume_lowtarif_perc	Percentage of consumption during the low tariff hours. From 10 p.m. to 7 a.m. and during weekends.
14- smartmeter_perc	Percentage of smartmeters in the zipcode ranges

Now that I have 64 files, 32 for electricity and 32 for gas, the next step is combining these files into 6 main data frames by merging on the columns: **zipcode_from & zipcode_to** for each network administrator and type of energy while adding 6 columns of each year as shown below

Final data frame name:

df + Network_admin + Energy_type

Final data frame structure:

Number of rows x **94 columns:**

6 fixed columns

1. net_manager
2. purchase_area
3. street
4. city
5. zipcode_from
6. zipcode_to

Columns added for each year: 88 (8 columns x 11 years – From 2008 to 2018)

1. num_connections_Year
2. delivery_perc_year
3. perc_of_active_connections_year
4. type_of_connection
5. type_conn_perc
6. annual_consume_year
7. annual_consume_lowtarif_perc_year
8. smartmeter_perc_year

Final data frame list:

Each data frame contain 10 years of data labeled for each column

df_Enexis_electricity
df_liander_electricity
df_Stedin_electricity

df_Enexis_Gas
df_liander_Gas
df_Stedin_Gas

Data Exploration & Visualization:

The following on from loading the data and preparing all the dataframe, now I have 6 data frame with the following shapes.

```

* Merged data for Electricity in enexis has final shape of: (90692, 86)
- - - - -
* Merged data for Electricity in liander has final shape of: (126360, 94)
- - - - -
* Merged data for Electricity in stedn has final shape of: (88948, 94)
- - - - -
* Merged data for Gas in enexis has final shape of: (63425, 86)
- - - - -
* Merged data for Gas in liander has final shape of: (97952, 94)
- - - - -
* Merged data for Gas in stedn has final shape of: (79621, 94)
- - - - -

```

Shape of final 6 dataframes

As observed above, data from Enexis network administrator has 86 columns instead of 94. This is due to the dataset missing values for the year 2008. For consistency I will be dropping 2008 observations.

Data Assessment & Cleaning:

Following an extensive data exploration process, the reoccurring theme of data assessment for all 6 data frames is as below – it is worth noting that in general the data is pretty clean due to the cleaning process done on the dataset before uploading to Kaggle.

Data assessment actions:

- Smart meters columns for Enexis Gas has not values for the year all years
- Missing values for 2009 type of **connection and type of connection** % for Enexis Electricity. To be completed from 2010 observation
- Some values for **Purchase_area** missing from Liander electricity
- Other minor missing values that have been dropped.
- Missing values have been reviewed and either filled or dropped.

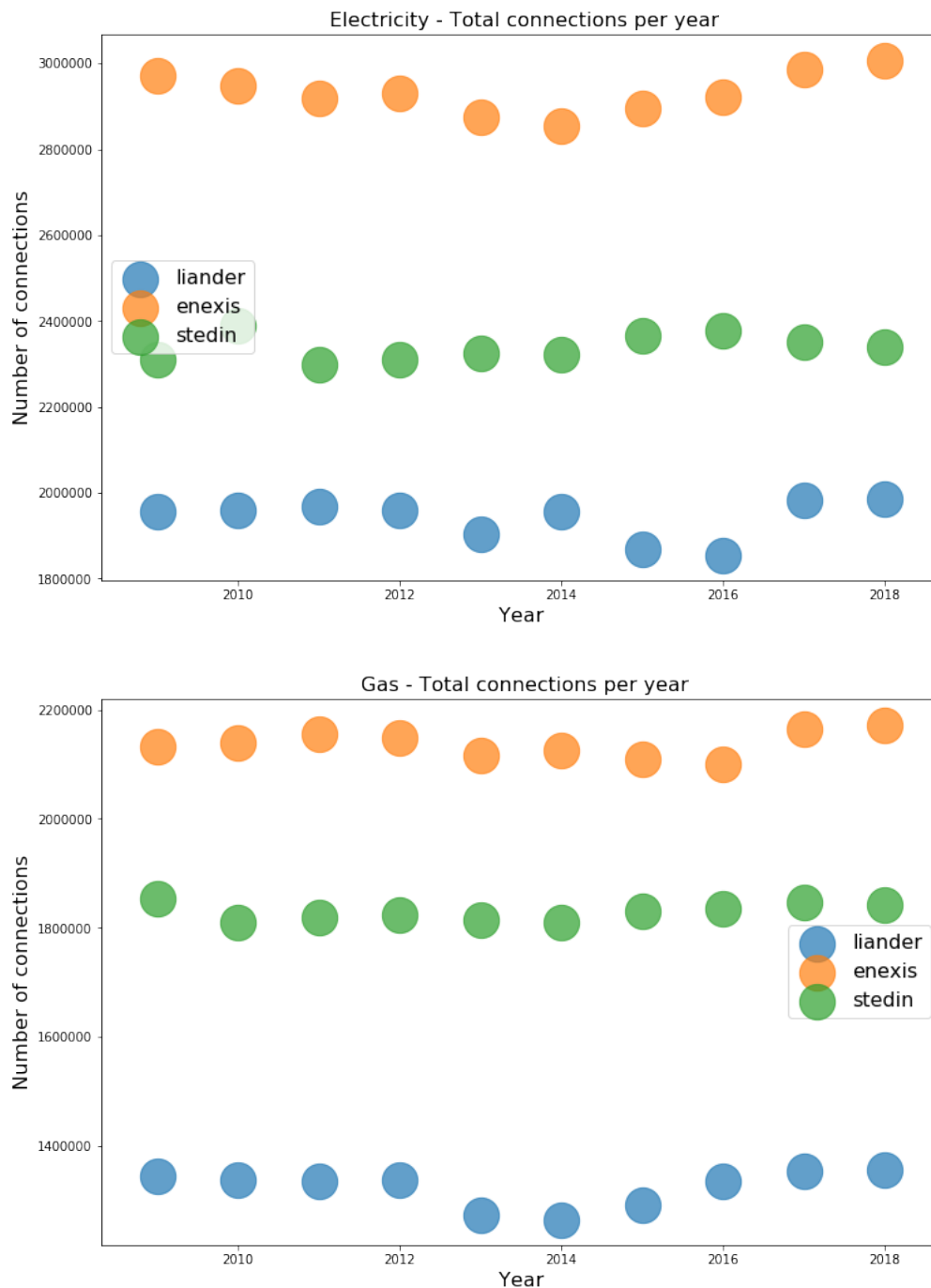
Visualization:

Data visualization and exploration have been completed in 2 sections:

- 1- Network Administrator Level
- 2- City Level


Network Administrator:

1. Firstly I explore Total number connections per year for each network administrator for both Energy Electricity and Gas:

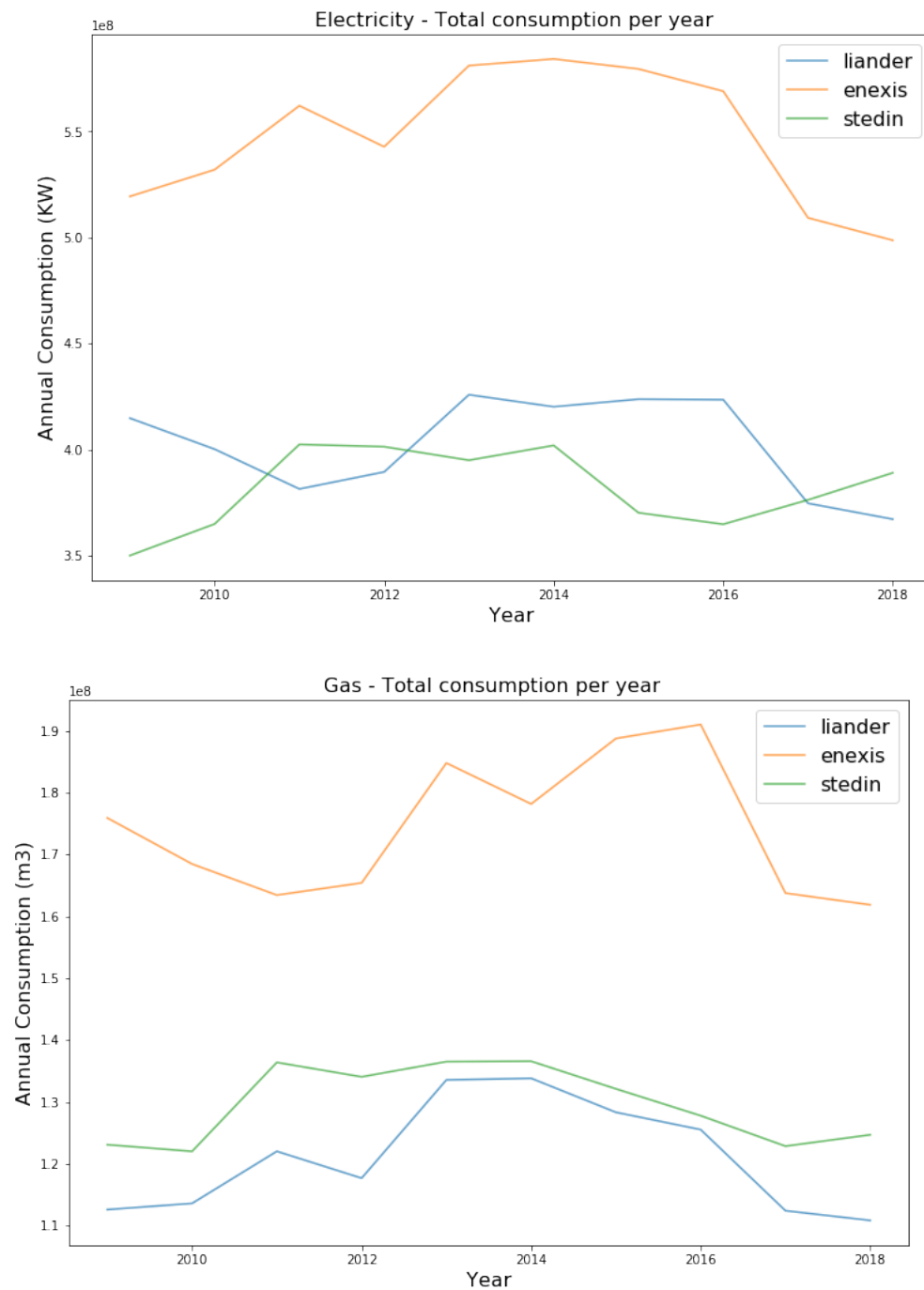


It can be observed from the above 2 graphs, that Enexis is by far the largest network administrator for total connections year on year. With the number of connections pretty much steady for the last 10 years.

The Dynamic score board

Number of Connections		
Enexis		
Liander		
Stedin		



- Secondly I will move to explore annual consumption per network admin, in order to establish if it follow same distribution as Number of connections seen prior.



Enexis still the biggest network administrator for both energy. Nevertheless, an interesting pattern emerge between Stedin and Liander. Even though they vary in term on number of

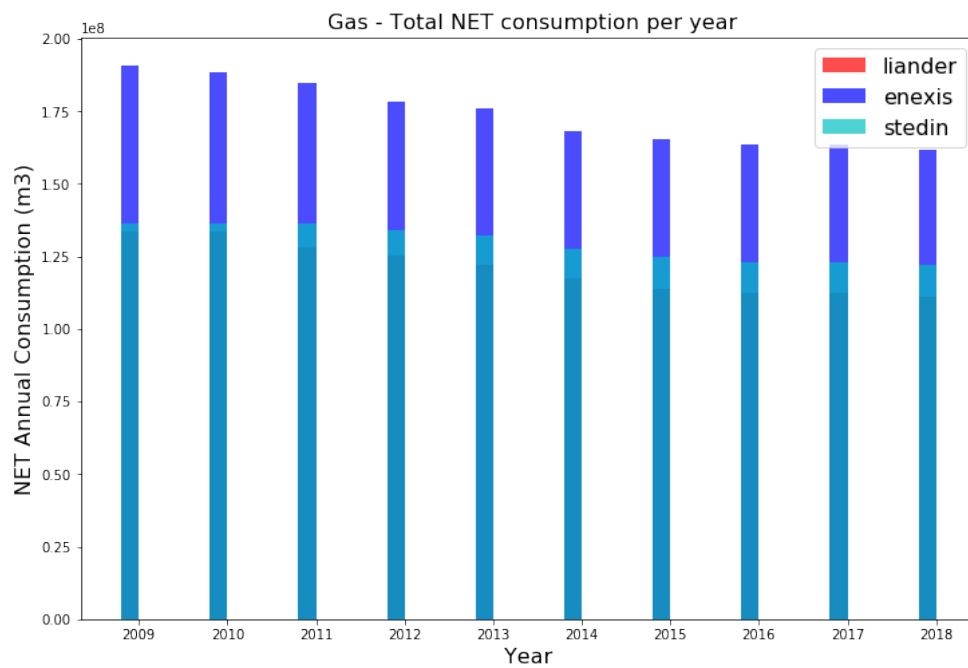
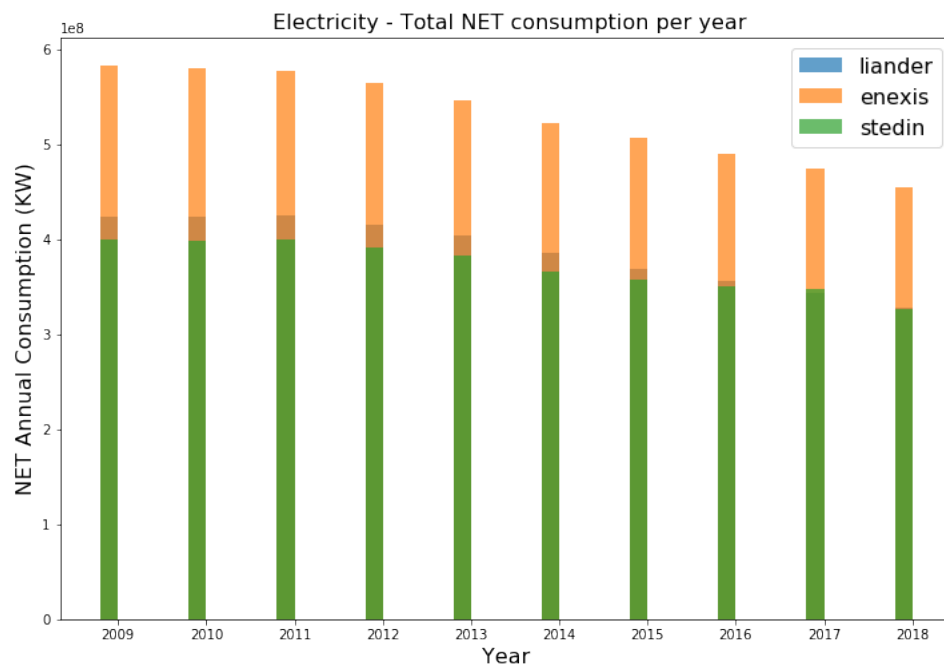
connections, the total energy consumed through their network is closely identical. This can be explained by Liander supplying more business customers or Stedin customers giving back to the network through their own production. Anyhow, I will explore this further down.

The Dynamic score board

	Number of Connections	Total Consumption
Enexis		
Liander		
Stedin		




3. Now, I will move into calculating net energy consumption. This is calculated as follows:

Consumption_per_year x delivery_percentage_per_year

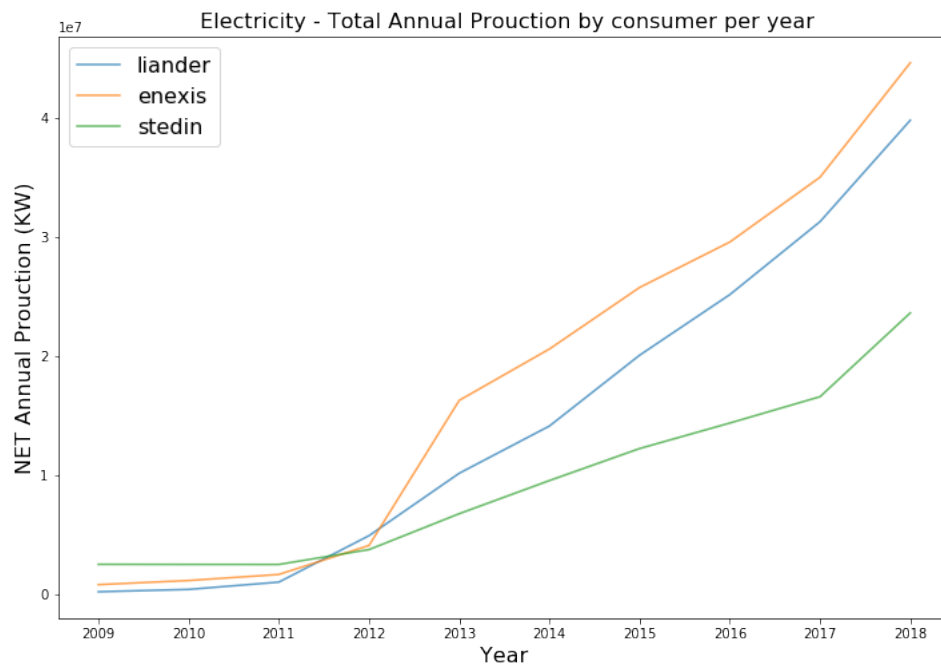


Net consumption is slightly less than total consumption , this is due to the self production being spread over the years, but Enexis still the clear leader in terms of net consumption.

The Dynamic score board





	Number of Connections	Total Consumption	Net Consumption
Enexis			
Liander			
Stedin			

4. Now that we have analyzed net consumption, I will move to explore the self production principal by the customers of each network administrator and review the trend over the years.

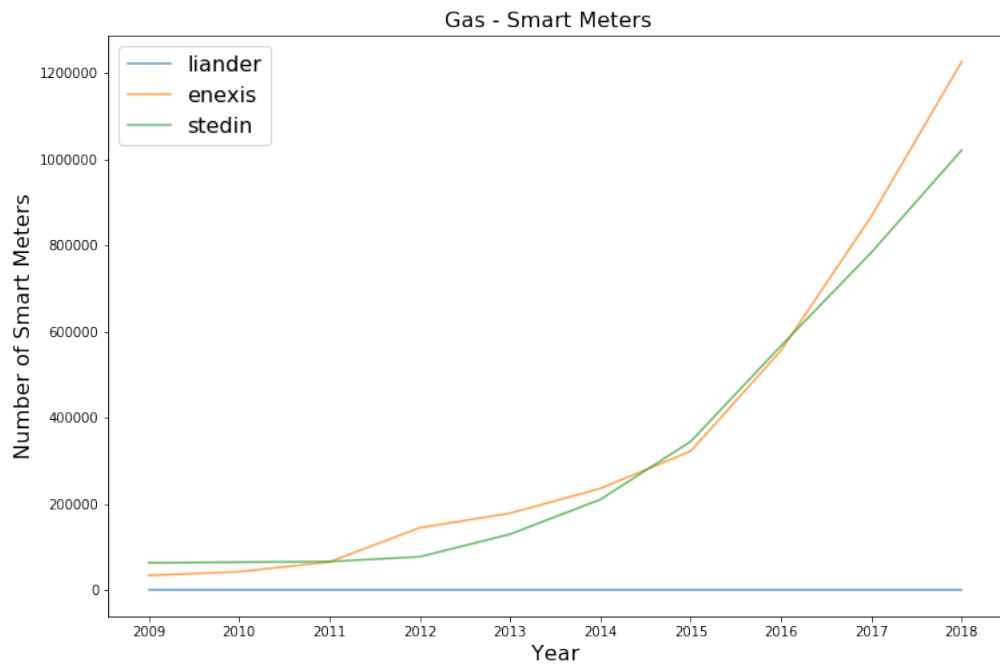
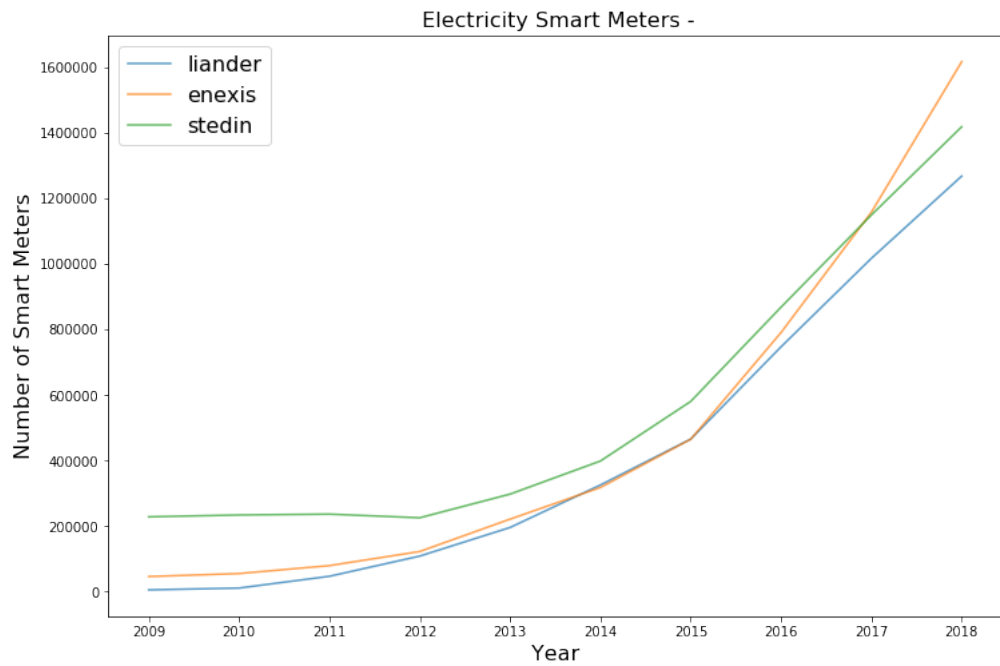


Very interesting results ! From 2012 in the electricity production consumers have steadily increasing their production of energy. This can due to the technology advancing from 2012 onward and people becoming more aware to produce clean energy domestically. Liander stand out in the annual production where their customers producing as much as Enexis!

The Dynamic score board







	Number of Connections	Total Consumption	Net Consumption	Production by cusotmers
Enexis				
Liander				
Stedin				

5- Smart meters penetration into the market is another interesting fact that I have explored. It is worth nothing that network Liander , has the gas smart meter missing from the dataset, hence the flat line at zero.



Liander and Stedin are clearly catching up to market leader Enexis, even though they have lower number of consumers. Their smart meter installing in their network is pretty high.

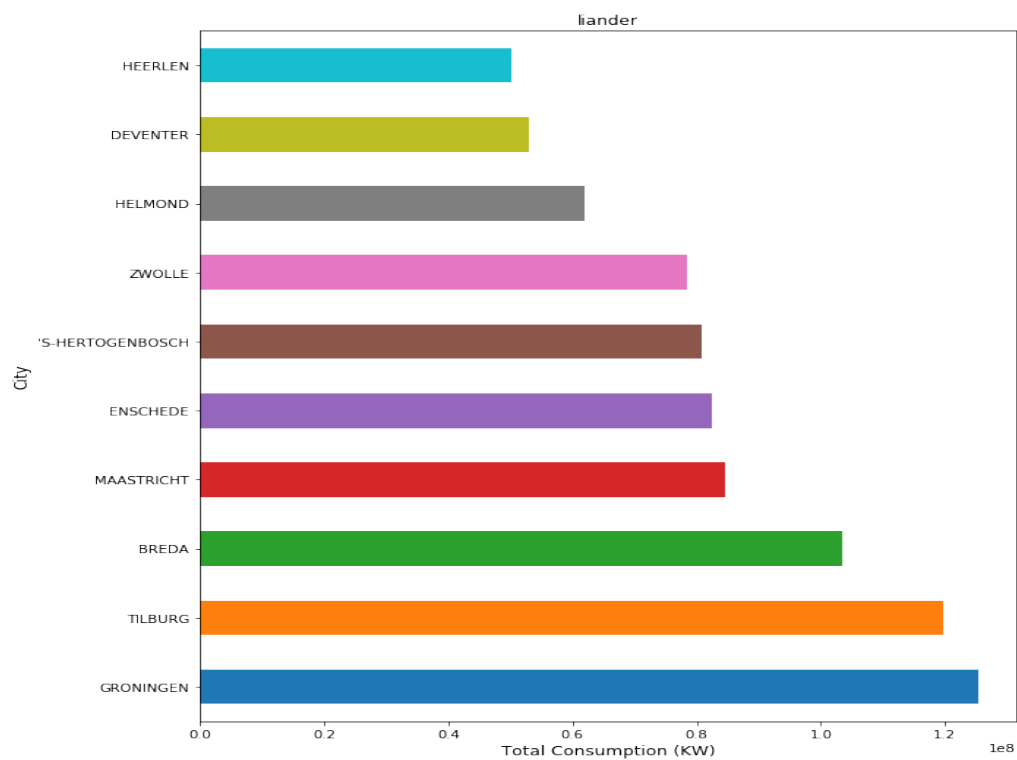
The Dynamic score board

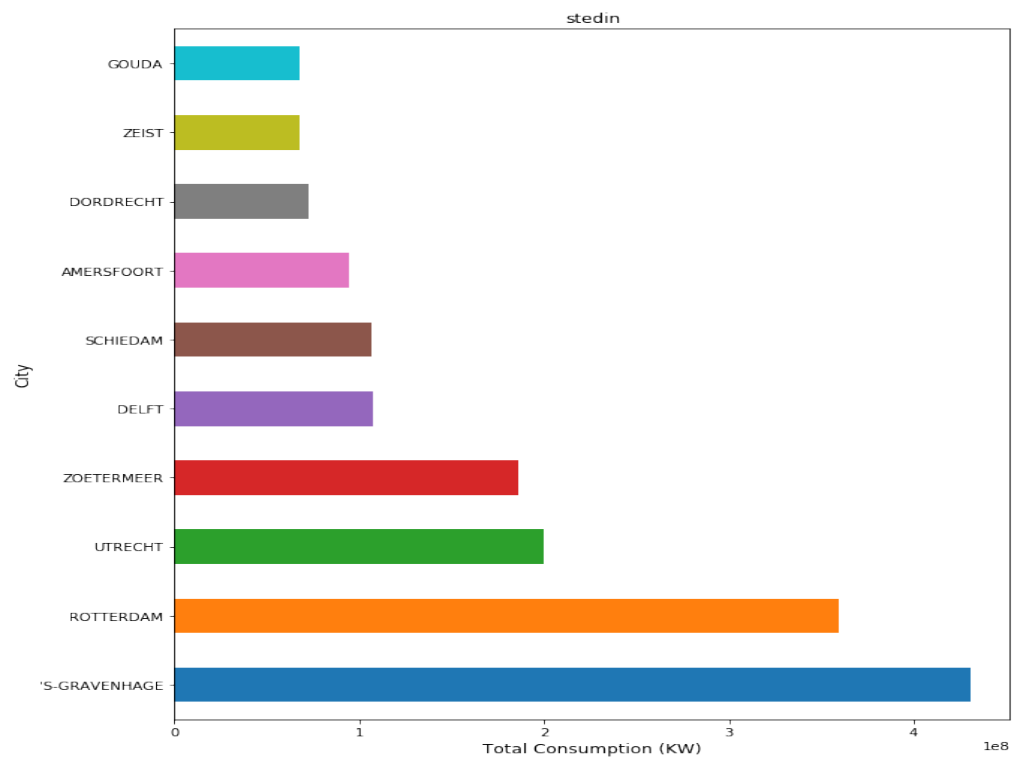
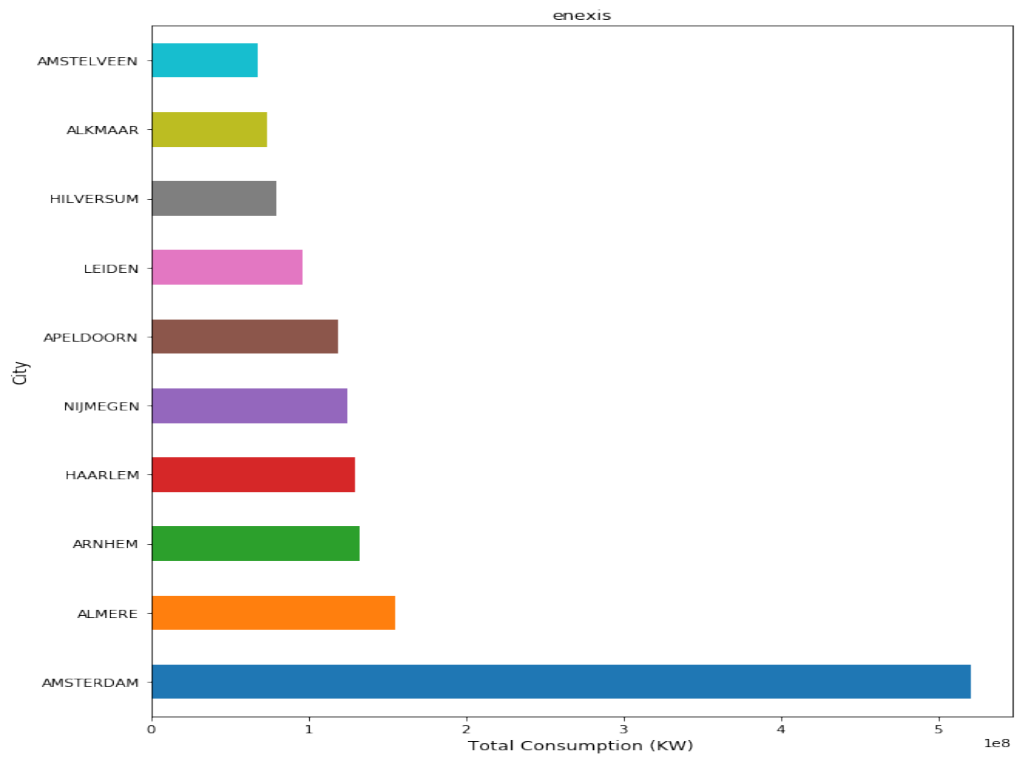
	Number of Connections	Total Consumption	Net Consumption	Production by customers	Smart Meters
Enexis					
Liander					
Stedin					

City Level:

- Initial review is for cities that consume the most energy in each of the network administrator.

Electricity:

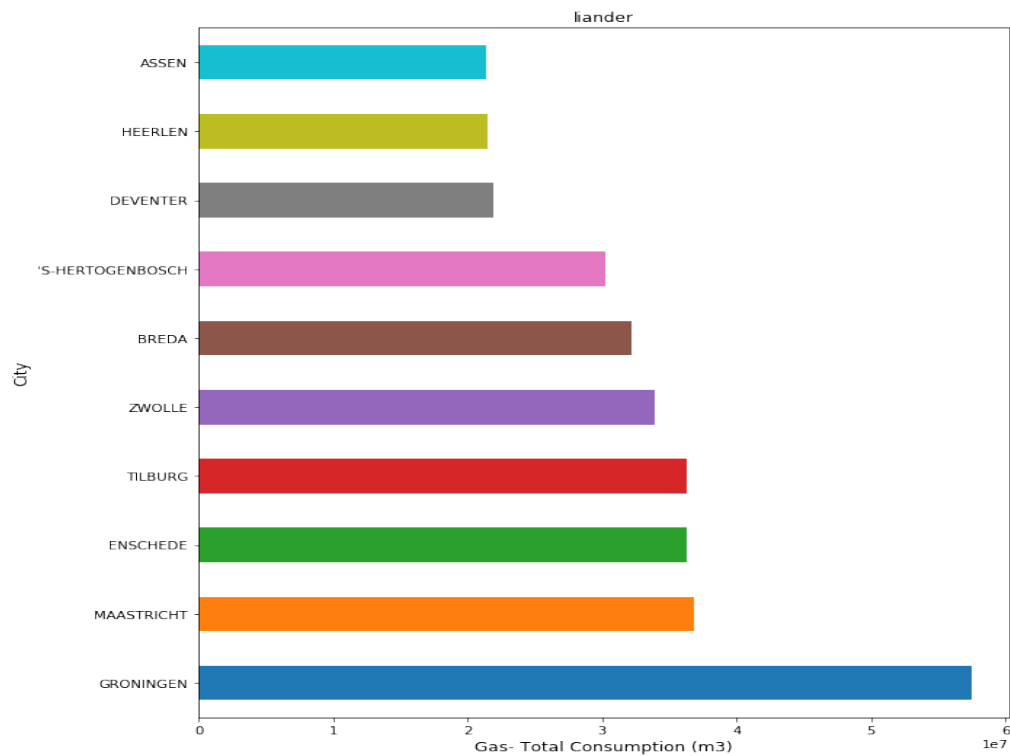


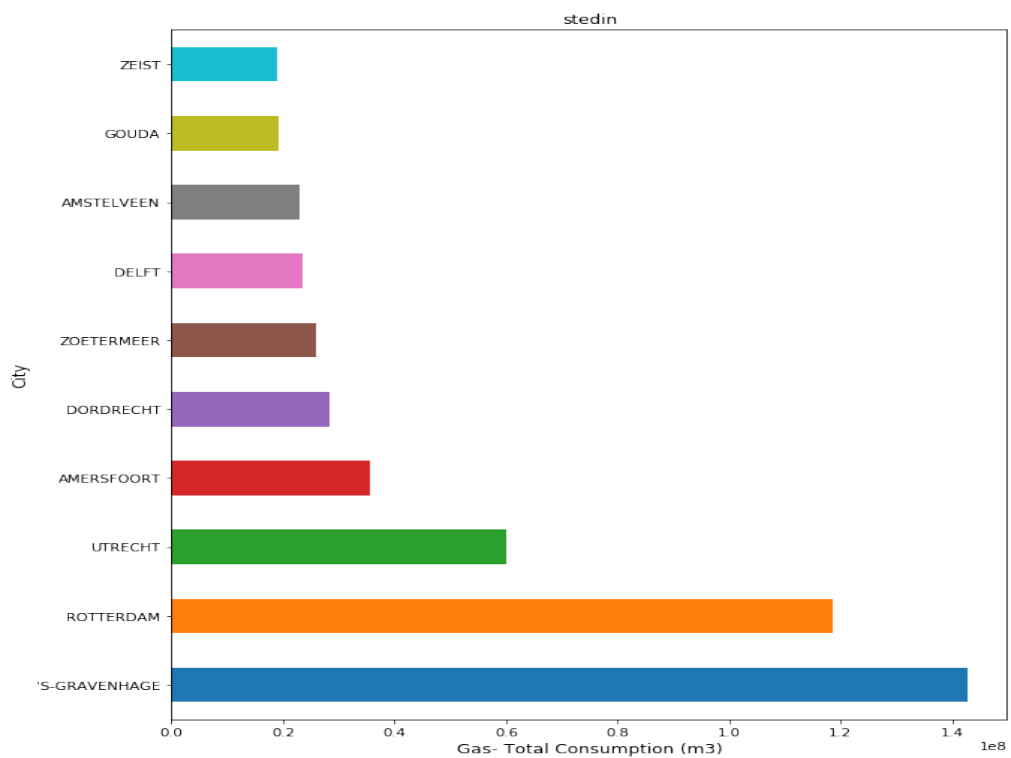
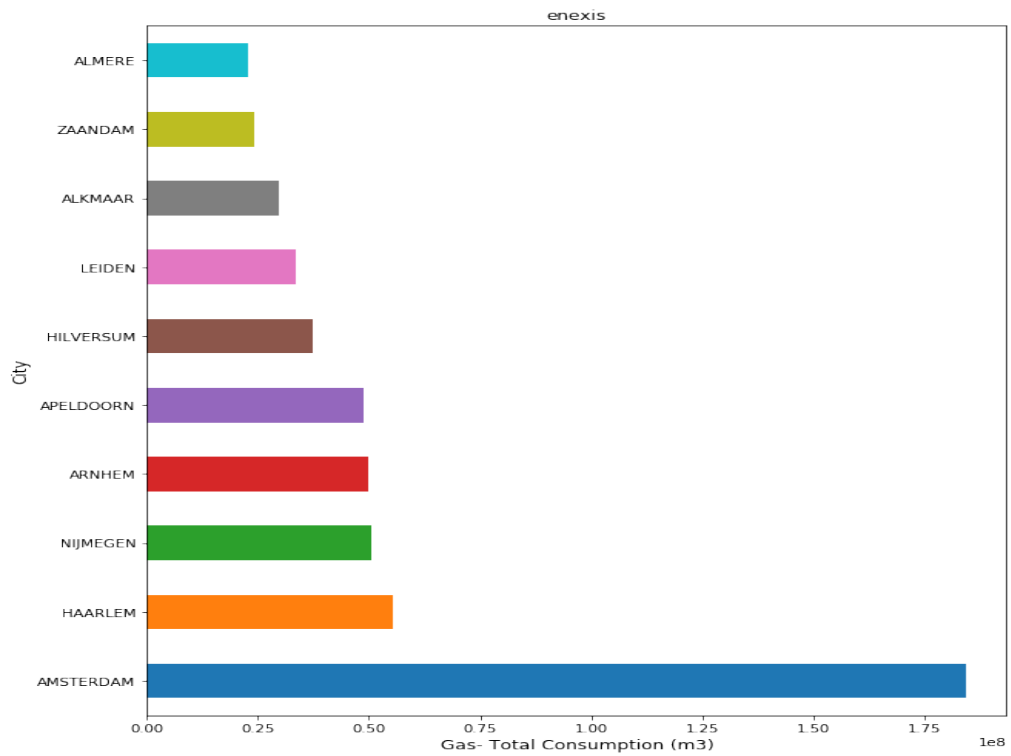


Big cities name pops up as expected for electricity consumption such as Amsterdam. It also can be noticed that Enerxis top consumption city , consume about 4 times as much as Liander top consumption city.

Amsterdam with population of about 800,000 consume 20% more than GravenHague that has a population of about 500,000.

Gas:

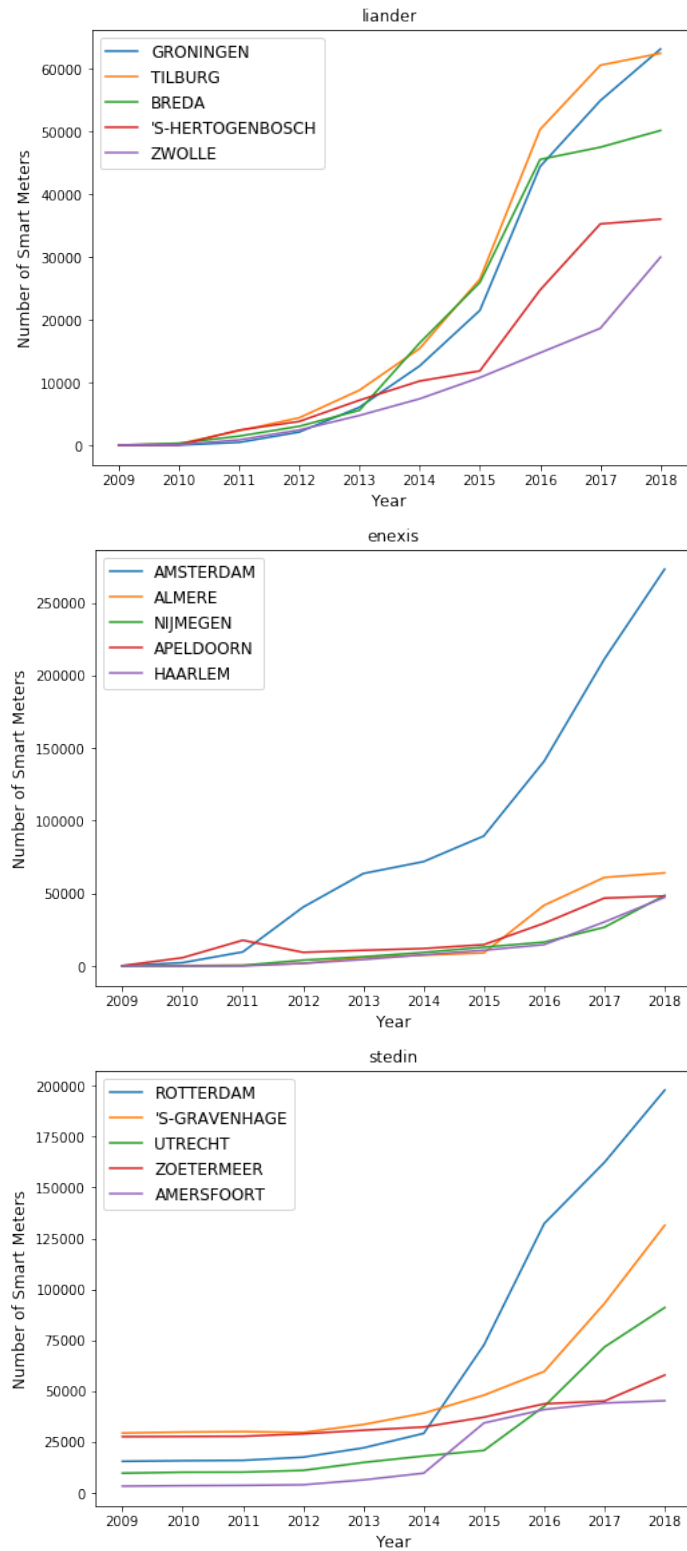




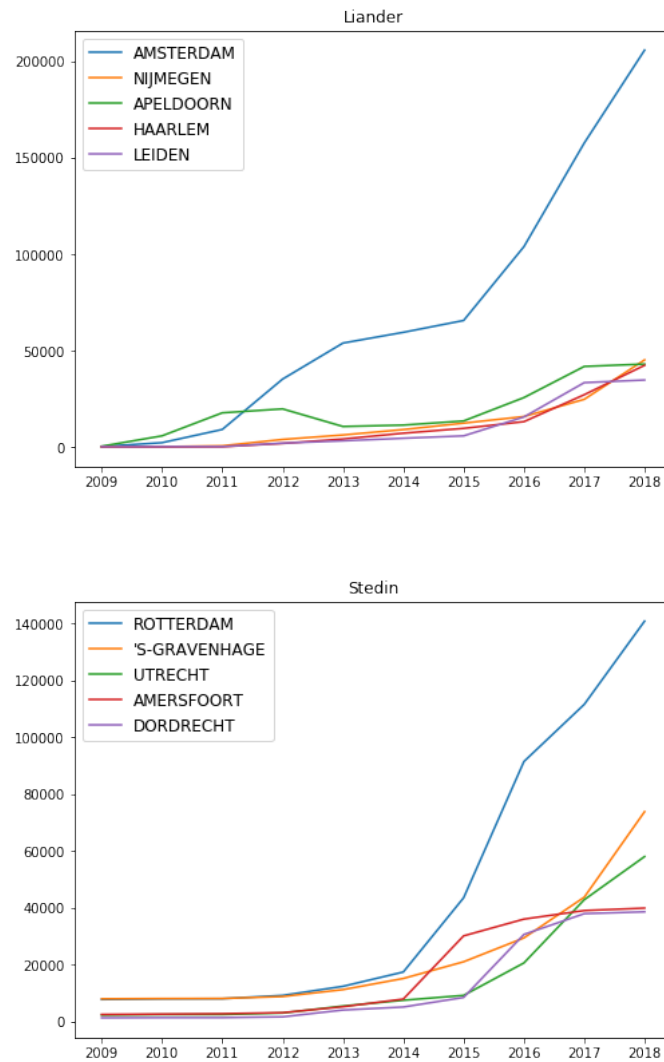
Big cities name also pops up as expected in similar way as Electricity.

- 2- Smart meter spread across the cities. In this section, I calculate the top 5 cities the have the most smart meters installed and what was their journey to reach this level of smart meters installation.

Electricity

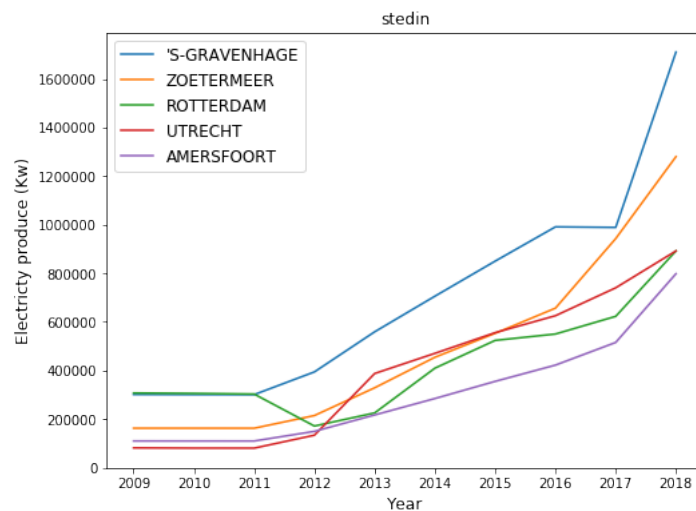
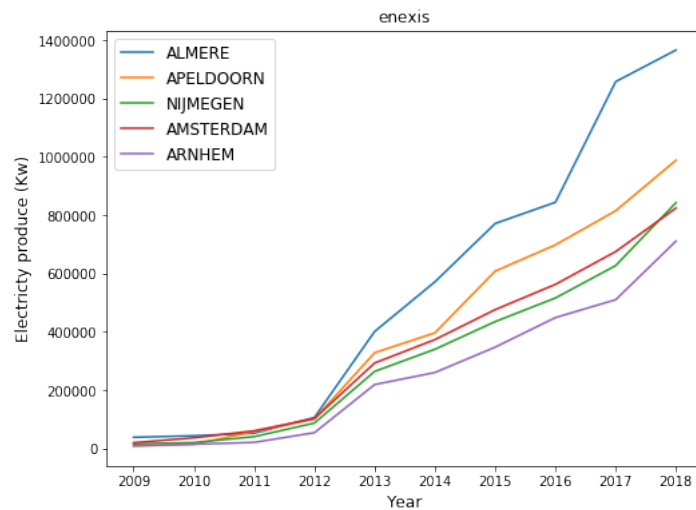
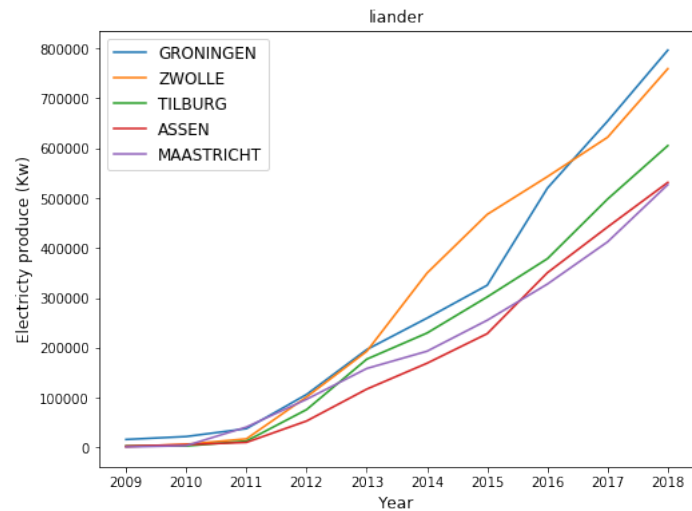


Gas



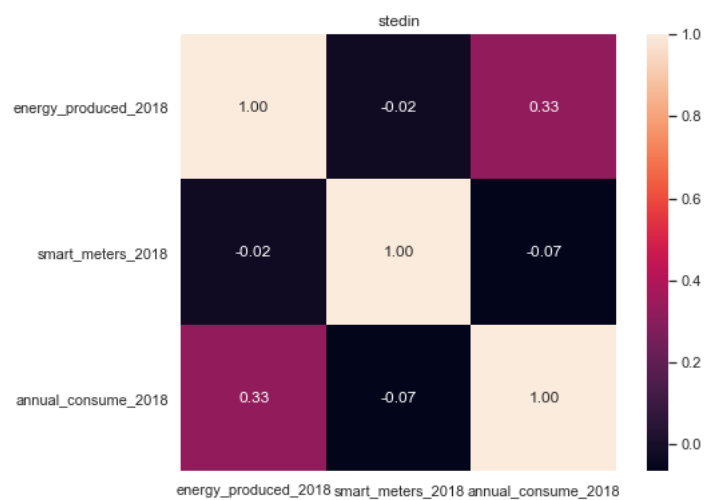
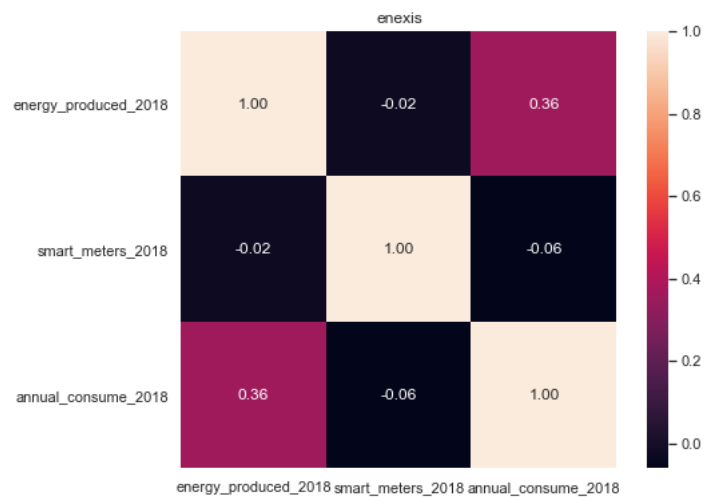
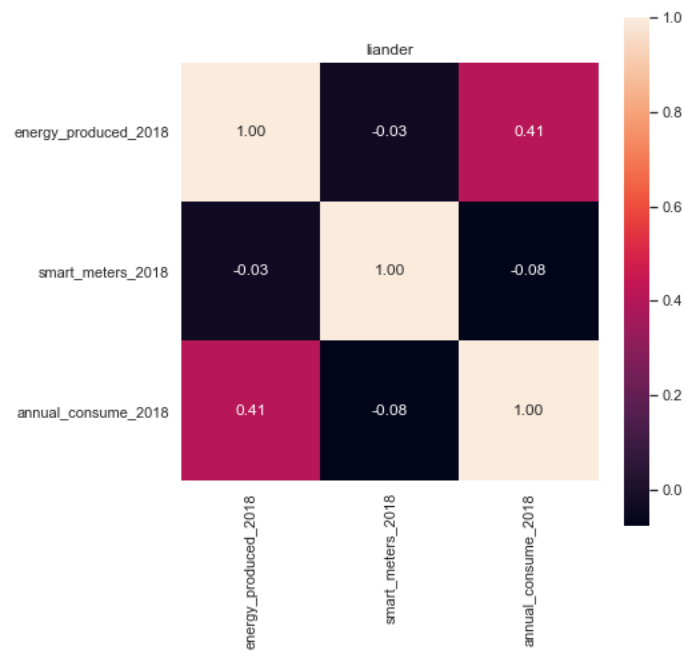
Rotterdam and Amsterdam are clear leaders in smart meter market penetration, this can be due to their respective high population compared to rest of Netherlands. Moreover, it is noticed that 2014 and 2015 is the years when smart meters started to rapidly spread in the country.

- 3- I will now look at the top 5 cities that have the most self production of Electricity, ie selling back to the grid. In addition I review their journey to this high level of self production throughout the past 10 years.



It is that 2012 is pivotal year to when cities thwarted to sell back to the grid. Moreover, the 2017 is also another year that representing a spike in Electricity produced domestically especially for the Stedin network administrator,

- 4- Finally, I explore data correlation between smart meter installation , energy produced by consumers and energy consumed for year 2018.



In all 3 network administrators I find a positive correlation (as expected) between energy produced by consumers and overall energy consumed for the year 2018. In addition a very small negative correlation is observed between smart meters and annual consumption.

Modeling:

Since the data is annualized and only available for 10 year, I decided to produce a dataframe combining the three network administrators dataframes for electricity while grouping by city. I plan to fit a model that will try to predict the dependent variable “Annual_consume_2018” by fitting the created independent variables for each city as below

Y_predict = annual_consume_2018

Features='num_connections_', 'smart_meters_', 'perc_of_active_connections_', 'delivery_perc'
(for each of the 10 years)

Dependent Var

Independent For Each year(10 years)

	annual_consume_2018	'num_connections'	'smart_meters_'	'perc_of_active_connections'	delivery_perc

A- Linear Regression Model

In Fitting a linear regression model, the below model evaluation were obtained. Size of test data was set to 33% and train data is for the rest 67%

Mean Absolute Error: 197.32523975387704

Mean Squared Error: 168267.50012869274

Root Mean Squared Error: 410.2042175900837

```
df_f['annual_consume_2018'].describe()
count      2282.000000
mean       5314.749781
std        2414.406613
min        1282.000000
25%        3949.016340
50%        4633.057059
75%        5848.175398
max        31623.410000
Name: annual_consume_2018, dtype: float64
```

Looking at the anual_conusme_2018 range, it is clear that the model is fitting the data well where RMSE is 410 compared to mean consumption for the year 2018 of 5,315.

It is also to be noted that additional Features such population, time of year, average family size can better

B- Random Forest

In Fitting a Random Forest model, the below model evaluation were obtained. Size of test data was also set to 33% and train data is for the rest 67%

Mean Absolute Error: 179.77019240202782
Mean Squared Error: 184572.52922049267
Root Mean Squared Error: 429.6190512773993

RandomForest is as good predictor model as Linear Regression with RMSE of 430 compared to 410 for LR.

Conclusion:

In this project I looked at 10 years worth of data for the Netherland 3 Energy network administrators, Enexis, Liander and Stedin. Enexis by far is the market leader when it comes to number of connections and total consumption per year for both Electricity and Gas.

Liander customer base are the country pioneers in term of self electricity production and feeding the reaming to the grid. It is worth further investigation for Liander strategy and they incentivizing their customers to self produce electricity.

When it come sto the number of smart meters installed nationally, all 3 network administrators are very closely equal in both energy Electricity and Gas , meaning that Liander and Stedin are doing great work in the spread of smart meters since there market share of total connection is below Enexis.

Zooming in to city level, my analysis show that as expected big cities Amsterdam, Hague and Gronningen are top energy consumers. In term of smart meters installation, Amsterdam big consumer base relate to the peak of smart meter installation where as other cities such as Rotterdam are leader in smart meter installation.

Self production in cities have taken off between 2012 and 2013, and maintained a steady increase. This may due to some legislation passed in the Netherland or technology such as solar panel has greatly improved. It is impossible to tell with this dataset.

I find a positive correlation (as expected) between energy produced by consumers and overall energy consumed for the year 2018.

Moving on to the modeling part of the report. In this part I applied 2 simple ML algorithms – Linear Regression and Random Forest. Both showed a very similar RMSE of about 400 while the mean of the predictive value is 5314